A THERMODYNAMIC AND STRUCTURAL DISSECTION OF COOPERATIVITY IN NATURAL AND DESIGNED TETRATRICOPEPTIDE REPEAT PROTEINS

by
Jacob D. Marold

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
July, 2015

ABSTRACT

A major goal in modern biophysics has been to thermodynamically characterize macromolecular systems to enable an energetic description of biological processes. Despite considerable effort, the thermodynamic nature of cooperativity in protein folding is not fully understood. The primary reason for this is due to the apparent “two-state” folding behavior at equilibrium, lacking intermediates. To grasp cooperativity in protein folding, one needs to thermodynamically quantify intermediates. Repeat proteins have proven to be excellent systems to thermodynamically describe these intermediates.

My work focuses on developing a thermodynamic description of protein folding cooperativity using two experimental systems of tetratricopeptide repeat proteins (TPRs/nPRs). nPRs consist of a repetitive n-residue motif, which forms antiparallel A- and B-helices. While our lab and others have had similar objectives on other repeat systems, my contributions have been 1) to develop and apply a statistical framework for analyzing heterogeneous systems, 2) to thermodynamically characterize units of structure smaller than whole repeats, 3) to ascribe structural bases to measured energetics, and 4) to understand mechanisms of stabilization by consensus design by studying a natural repeat protein system.

Consensus ankyrin and leucine rich repeat proteins are characterized by very unfavorable intrinsic folding free energies and strong interfacial interactions. In contrast, isolated c34PRs have a K_{eq} ~1 for folding, while interactions between helices are more modest. To determine the molecular origins of cooperativity in c34PRs, in Chapter 2, I develop and present a single helix heteropolymeric Ising model capable of resolving energies of half repeat units in nPR systems. I applied this model to consensus TPRs (c34PRs), and quantified energetics of single α-helices, as well as inter- and intra-repeat interfaces. While c34PR helices have different intrinsic energies, inter- and intra-repeat interfaces are similar in energy, despite structural differences.

In Chapters 3 and 4, I studied a naturally occurring 42PR with a longer sequence motif. I solved the X-ray crystal structure of five tandem 42PRs, and determined the longer sequence motif to result in helical extensions of the canonical helices of 34PRs. I quantified folding cooperativity in this system by using nearest-neighbor models in Chapter 4. 42PRs are more cooperative than 34PRs, due to increased magnitudes of both intrinsic and interfacial energies. Point substitutions suggest a single hydrogen bond in *Pa* 42PRs to contribute significantly to interfacial stability.

PRIMARY READER: DR. DOUG BARRICK

SECONDARY READERS: DR. JULIETTE T.J. LECOMTE and DR. VINCENT J. HILSER

THESIS COMMITTEE: DR. DOUG BARRICK, DR. VINCENT J. HILSER, DR. BERTRAND GARCIA-MORENO, DR. JULIETTE T.J. LECOMTE, AND DR. ELIJAH ROBERTS

For my mother and grandmother, the strongest women I will ever know

ACKNOWLEDGEMENTS

I would like to start by thanking my advisor, Dr. Doug Barrick. I can still remember my first interaction with him at the end of my recruitment/interview weekend. Although we did not speak to one another until the last hour, he shook my hand and said, “Jake, we never had a chance to talk this weekend, but I am hoping we can find some time to do that soon.” Throughout my time at Hopkins, he has continued to impress me by his unwavering enthusiasm for science, his devotion to all students, and his superb critical thinking ability. I have learned so much from him about all areas of life, and I will forever be grateful for the time he has invested in my education, as well as ketchup and mustard. I will miss his painted fingernails, sport sunglasses, baggy clothes, analogies, attention to detail, and chalk board prowess. I will not miss his cherry-tone figures (there are many more colors in this world), or stories about his broad sword (that hobby is just weird).

I would like to thank my thesis committee members, Drs. Juliette Lecomte, Vincent Hilser, Bertrand Garcia-Moreno E., Mario Amzel, and Elijah Roberts. They have all contributed significantly to my development by challenging my science, and always pushing me to reach for something greater. I would like to especially thank Juliette for attending every one of my committee meetings, and for her extra efforts. She has been a second advisor to me. Despite the fact I did not end up doing much NMR, in many ways she deserves as much credit as Doug for contributing to my personal and professional development.

I would like to thank Dr. Ananya Majumdar for all of his advice, support, and friendship over the past six years. I am very happy to have him in my life, and I know we will remain in close contact.

With so many past and present members of the Barrick Lab, it is hard to give everyone detailed thanks, but they deserve it in their own ways. All of them deserve my thanks for dealing with my messiness and my wild (yet adorably sweet) dog Raven. Dr. Thuy Dao is my mentor and sister. We have learned a lot from each other about both science and life, and I am happy to have met such a wonderful person to share part of my graduate career with. Although I only knew Dr. Andrea Allgood (Carter) for a short while, I have never forgotten all she taught me about effective communication. Dr. Scott Johnson taught me many things about being practical in both science and life. Dr. Tural Aksel taught me to be more considerate, and Dr. Eva Cunha taught me to enjoy life both in and out of the lab, and to realize the bigger picture. Kate Sherry has taught me to be more careful about giving her opportunities to make fun of me (lightheartedly), and we have had many good conversations about science and life over beer at One World, or Thanksgiving dinners with Moscow mules. I am excited to continue those conversations as we transition into our next paths. Kevin Sforza and Sean Klein will both go far in this world if they continue to strive as hard as they have been. Christine Hatem is very intelligent, and the sweetest person I will ever know. She has taught me to be resilient and to always push forward. Although I have my big (little) sister Thuy, I was fortunate to have been gifted a little sister,

Katie Geiger (will never be Schuller to me). Katie and I were bay mates and have benefited from each other's different perspectives. We have learned a lot about science and life together, and even though I told her I hated when she looked over my shoulder to see what I was coding/working on, I secretly enjoyed it so that I could teach her about it.

I would like to give a special thanks to Dr. Katie Tripp. In many ways we are soulmates, and I love that I have had the opportunity to interact with such a beautiful person. We have a unique understanding of each other and can be both critical and compassionate in many contexts. She has always pushed me to be a better person, and has offered support when I have needed it most. Although I was good friends with Brian Tripp (her cousin) in high school, she has most certainly surpassed him, and will forever be my favorite Tripp.

PMB is filled with so many wonderful people, and I have benefitted from all of them in different ways. Specifically, I would like to thank Drs. J.D. Schonhoft, Helen Jun, Aaron Robinson, Matt Preimesberger, Carla Coltharp, Matt Pond, Robert Trachman, Andrew Buller, and Jackson Buss. I would like to especially thank Peregrine Bell-Up. He is very intelligent, and a special person in my life. I would also like to thank Robin Thottungal, Jesse Yoder, and Hesam Motlagh.

I thank all members of the office staff – both past and present. I would be lost without Jerry Levin, Jess Appel, Ken Rutledge, Lexie Ebert. I thank Nicole Goode for keeping me on top of graduation timelines – she has done an excellent job. Ranice Crosby is one of my dearest friends. I would be lost in both life and in

my biophysics progression without her. She has taught me many things (including how to re-learn how to fix a sink) in life, and I will always love her from the bottom of my heart.

I would like to thank my family for their support through these years. Even though they might have been annoyed with my science talk at times during holidays, I am grateful to them for listening. My mother deserves more thanks than anyone, and she is the strongest person I know. I would have gone astray long ago were it not for her guidance.

Lastly, I would like to thank Alison Neal. She is a very special person in my life, and I would not have been able to succeed without her. She is the most passionate and caring person I know, and will make a fantastic Dr. in the near future. She deserves as much credit for this thesis as I do. Of course, I have to thank my dog Raven Taylor Marold. She was my companion at night, and I would not have been able to get as much work done without her keeping me happy at 4 AM when we would still be working in lab.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

CHAPTER 1

Introduction

One of the most beautiful reactions in biology is the process by which proteins acquire their three-dimensional folds. Spontaneously without the input of energy, linear protein chains of an overwhelming variety of sequences form both specific and general intra-molecular interactions that govern the structure of the native state. This process can be seen from many perspectives, but there has been wide general interest in the kinetics and thermodynamics of protein folding. The inception of the thermodynamic hypothesis in protein folding was arguably conceived during early studies of ribonuclease (RNase) (Sela et al., 1957; Anfinsen, 1973), where simple combinatorics of disulfide linkages allowed researchers to discover that only one correct set led to enzymatically active protein. The field of protein folding has had a rich and fruitful history (Baldwin, 2007; Dill, 1985; Englander et al., 2007; Honig, 1999; Sosnick and Barrick, 2011), and it is outside of the scope of this dissertation to cite every noteworthy accomplishment. However, despite the concerted efforts of scientists over nearly five decades, questions remain regarding the nature of cooperativity in folding, which will be the center point of this work.

1.1 Protein diversity and function

Proteins are at the heart of cellular activity and represent some of the most diverse biomolecules. From classic examples of folding models with minimal hydrophobic cores like the 35-residue Villin headpiece subdomain (McKnight et al., 1996) to the massive 33,000 residue titin (Bang et al., 2001), proteins are chemically diverse simply from the size of their sequence space and chain lengths. Moreover, proteins have an additional layer of complexity, as they are capable of remarkable spontaneous self-organization through intra and inter-molecular interactions. These characteristics allow them to adopt many conformations and interact with a variety of other macromolecules to accomplish cellular tasks (Alberts et al., 2010; Berg et al., 2007).

In many ways, protein complexity and diversity can be viewed as a biophysical extension of the central dogma of biology, which was originally developed to explain the transmission of information from DNA (Crick, 1970). Extending from this description, information transfer can take many forms other than the chemical composition of the macromolecule, and a "residue-by-residue code." For example, while mRNA splicing events result in distinct protein chains after translation, the way in which these chains encode inter- and intra-molecular interaction represents another form of protein diversity. Proteins undergo conformational changes, bind ligand(s), and can self-associate. These features can be influenced by

environmental variables such as osmolytes, temperature, pH, macromolecule concentrations, etc. Therefore, it is the interplay between the encoded chemical information and external factors, which determines how originally encoded DNA “information” results in functional outcomes (Figure 1.1).

It is of great interest to understand how factors determine the conformational and functional landscapes of proteins. A quantitative description of protein folding and protein-protein or protein ligand interactions enables remarkable predictive power for carrying out experimental, physiological, or environmental applications. Two powerful frameworks—thermodynamics and kinetics—have comprised the backbone of understanding biomolecular processes. Thermodynamic measurements provide a scale for relative strengths of intra- (conformational stability) and inter-molecular (protein-protein/protein-ligand) interactions, while kinetic descriptions yield information on time-scales of biological processes (Van Holde, 2006). Biophysical characterization of how different variables shape kinetics and thermodynamics of processes provides detailed mechanistic insight into how function is determined.

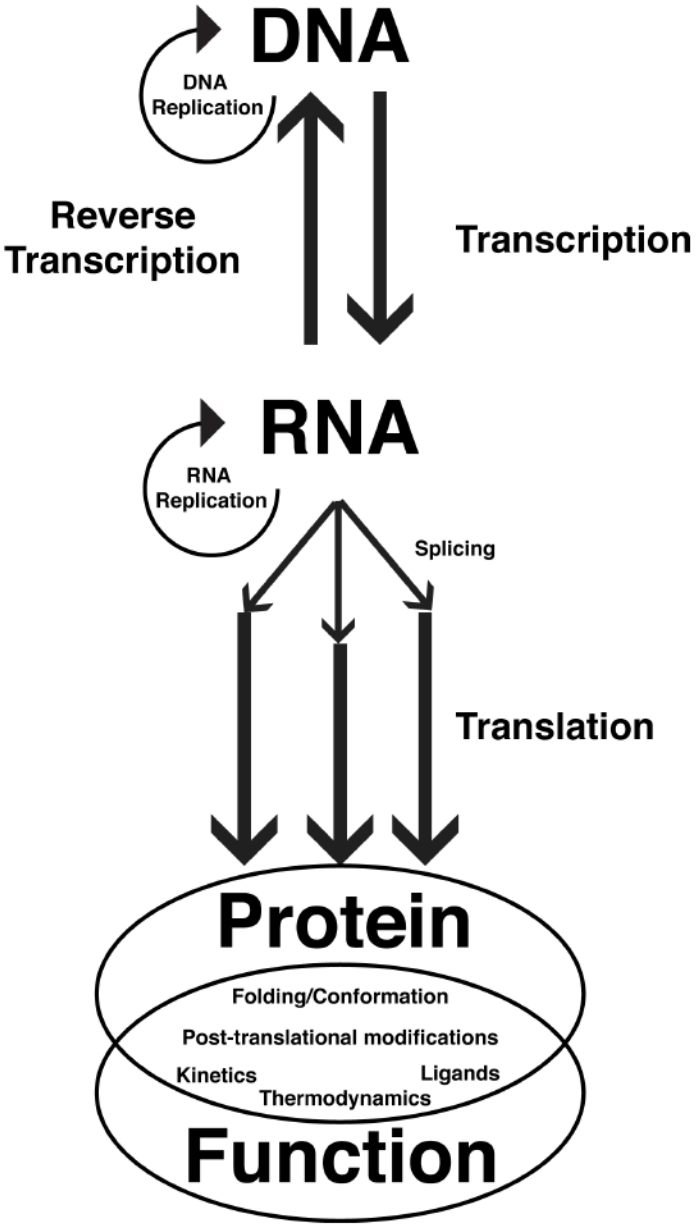


Figure 1.1. Protein diversity viewed as an extension of the central dogma. Arrows indicate directional information flow. The Venn diagram highlights factors influencing protein function. It must be emphasized that these factors display contextual dependencies, which can be governed by concentration(s), pH, solvent, osmolytes, temperature, etc.

1.2 Cooperativity in Macromolecular and Chemical Systems

Cooperativity is a defining feature of many biological and chemical phenomena (Cui and Karplus, 2008; Englander et al., 2002; Koshland Jr et al., 1966; Liu et al., 2007; Qian, 2012; Sharp, 2001a; Shea and Ackers, 1985; Zimm and Bragg, 1959; Monod et al., 1965). Because of this, cooperativity has become loosely defined. Despite this, there are universal features of all cooperative processes, and this section will highlight these, and connect each to protein folding.

Water, phase transitions, and critical phenomena

A crude description of cooperativity is to envision it as the process by which reactions proceed in an "all-or-none" fashion—a lack of intermediate states at equilibrium. Physical phase transitions provide excellent conceptualizations of cooperativity (Stanley, 1987), and even the behavior of water can serve as a rich source of physical intuition. Extensive hydrogen bonding networks determine relative H-bond strength between H_2O molecules, and energetically connect physically distant molecules (Sharp, 2001; Gerstein and Levitt, 1998; Eisenberg and Kauzmann, 1969).

Critical points on phase diagrams describe two distinct states with different physical properties (such as density in the melting of ice) in

equilibrium that can be cooperatively shifted (Dill, 2010). A universal feature of cooperative reactions is that they have high heat capacity (C_p) at transitions. This hallmark also extends into biological macromolecules (such as in conformational transitions), although C_p changes are generally lower (Van Holde, 2006). The melting of lipid vesicles, for example, displays a sharp heat capacity change measured by differential scanning calorimetry. However, heat capacity changes alone cannot determine the *extent* of cooperativity.

Heat capacity as a measure of cooperativity in biological systems

There are many biophysical methods available to measure ΔC_p in reactions including direct model-independent approaches such as differential scanning calorimetry (DSC) (Freire, 1995; Sturtevant, 1987; Hühne et al., 1996; Bruylants, 2005), or by studying a temperature dependence of the equilibrium constant (K_{eq}) and applying van't Hoff or Gibbs-Helmholtz relationships (LiCata and Liu, 2011). In addition to providing numerical values of reaction enthalpies, information from calorimetric studies provide a wealth of mechanistic insight. Positive ΔC_p is associated with the hydrophobic effect and the solvation of non-polar chemical groups, while a negative ΔC_p is associated with exposure of polar groups. Typically, the unfolding of small globular proteins has displays

positive ΔC_p, consistent with the exposure of hydrophobic groups to solvent (Prabhu and Sharp, 2005).

One powerful methodology has been to use van't Hoff relationships in tandem with DSC. Since van't Hoff assumes "two-state" behavior, a comparison of measured ΔH's obtained from DSC and van't Hoff relationships can be used to determine whether a reaction proceeds in a highly cooperative manner. In protein folding, this is often used and discussed as test of the "two state" hypothesis (Freire and Biltonen, 1978; Privalov and Dragan, 2007; Zhou et al., 1999; Liu and Sturtevant, 1995, Liu and Sturtevant, 1997). However, the result from this test does not yield mechanistic insight into cooperativity.

Early observations of structural changes in hemoglobin upon cooperative oxygen binding (Perutz, 1970) have formed the foundation of cooperativity in ligand binding events, or multi-molecular reactions involving conformational changes and allostery. However, unimolecular reactions can also show varying degrees of cooperativity. Cooperativity does not have to be "yes" or "no", but can present itself as a continuum, yet there is no universal range separating cooperative and non-cooperative processes. My goal has been to provide a quantitative measure of cooperativity in protein folding, which provides a basis to map

the cooperativity range. To motivate the discussion of the ranges of cooperativity, the helix-coil transition provides an excellent framework.

Helix-coil theory

Helix-coil transition theory can provide insight into degrees of cooperativity. In the interest of brevity, I will not go into mathematical derivation of the models, but instead refer to Figure 1.2, where I have summarized expressions for partition functions for some historical models and how to use them to calculate fractional helix populations. For a more detailed description of these models, their applications, and their advancements I refer to (Doig, 2002; Doig et al., 2001; Zimm and Bragg, 1959; Dill, 2010; Cantor and Schimmel, 1980).

In the simplest limit of a noncooperative system (every residue is independent), a significant fraction of intermediate states is predicted, whereas for the two-state approximation, there are only two states observed, and they are separated by an energy barrier at the midpoint of the transition (Dill, 2010). The Zimm Bragg treatment is of particular interest to the work in this thesis, as the “s” term describes the equilibrium between a folded state and a reference state, and σ represents the barrier required to initiate helix formation in the absence of neighboring residues in a helical conformation. This representation is analogous to the

approaches used in generating partition functions for systems of repeat proteins for analysis using Ising models.

While simplistic, the expressions in Figure 1.2 are powerful to describe degrees of cooperativity in systems when their application permits. The matrix representation of the partition function was taken from (Dill, 2010) and will be the formalism used in Chapters 2 and 4.

	q	f_H
Noncooperative model	$(1 + s)^n$	$\frac{1}{(1 + s)^n}$
Two-state model	$1 + s^n$	$\frac{s^n}{(1 + s)^n}$
Zimm-Bragg	$[1 \quad \sigma s] \begin{bmatrix} 1 & \sigma s \\ 1 & s \end{bmatrix}^{n-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\frac{1}{n} \frac{\partial \ln q}{\partial \ln s}$

Figure 1.2. Overview of common helix-coil transition models. s and σ represent statistical weight terms, q is the partition function, f_H is a function to relate to the fraction of helix, and n corresponds to the number of “residues” in each model.

1.3 Repeat proteins as a tool for understanding cooperativity in folding

Protein folding has been at the heart of biophysics since the first structure determinations of myoglobin and hemoglobin (Fersht, 2008). Since then, collective efforts have resulted in extremely detailed information regarding the folding process (Barrick, 2009; Sosnick and

Barrick, 2011). Much of our knowledge progression has been due to experimental advancements. X-ray crystallography allows us to view intricate atomic details of static structures (Wilkins, 2013; Matthews, 2012), nuclear magnetic resonance (NMR) experiments residue-level resolution of conformational changes and protein-protein interactions in solution (Englander and Mayne, 2014), and single molecule tweezers now allow us to bypass solution ensemble descriptions (Jagannathan and Marqusee, 2013; Moffitt et al., 2008) and experimental bottlenecks such as aggregation. In addition, molecular dynamics simulations can now be run for milliseconds to visualize details about folding pathways (Lindorff-Larsen et al., 2011).

While molecular dynamics simulations provide insight, it is still extremely difficult to experimentally characterize protein folding cooperativity in detail. Recent advancements in hydrogen-deuterium exchange NMR and mass spectrometry (MS) have helped shape our understanding of cooperativity, but they only paint part of the picture, as their thermodynamic interpretation is often auxiliary (Bai and Englander, 1996; Englander et al., 2002, 2007; Hu et al., 2013). What factors govern cooperativity in protein folding? How universal are they? What are the underlying principles?

Factors limiting the understanding of protein folding cooperativity in globular proteins

In order to quantify cooperativity in folding, intermediate states and their structures must be independently studied. Some proteins have displayed multistate unfolding and can be analyzed appropriately to determine stabilities of intermediates and their populations (Barrick and Baldwin, 1993; Kuznetsova et al., 2002; Ionescu et al., 2000). However, the high degree of cooperativity in protein folding generally suppresses intermediate states.

While there are universal physio-chemical features regarding energetics of protein folding, our ability to dissect these energetics is limited (Baldwin, 2007; Dill, 1985; Chan et al., 1995). Globular proteins can have complex topologies, and can form native contacts between residues distant in sequence. This can lead to highly interconnected structures, and make it difficult to determine energetic contributions of structural units (Pascarella and Argos, 1992; Shortle and Sondek, 1995). There are even proteins which form knots, and display complicated folding behavior (Mallam and Jackson, 2007; Mallam et al., 2008). In contrast, repeat proteins are not subject to many of the structural engineering constraints imposed on globular proteins, and therefore provide an avenue to explore contributions of structural elements to folding cooperativity.

Repeat proteins as tools for understanding folding cooperativity and connection to multidomain proteins

Linear repeat proteins are composed of small structural motifs which stack in tandem (Kajava, 2001; Kloss et al., 2008). Individual motifs are structurally diverse, yet they share the common feature of linearity (Figure 1.2). One useful comparison is to connect repeat proteins to multidomain proteins (MDPs). MDPs are formed from arrangements of large independently folding units on a single polypeptide chain (Capp et al., 2014; Vogel et al., 2004). Figure 1.2D shows an example of the structure of the Ig domains of the MDP titin. When viewed from this perspective, repeat proteins are linearized microscopic versions of MDPs.

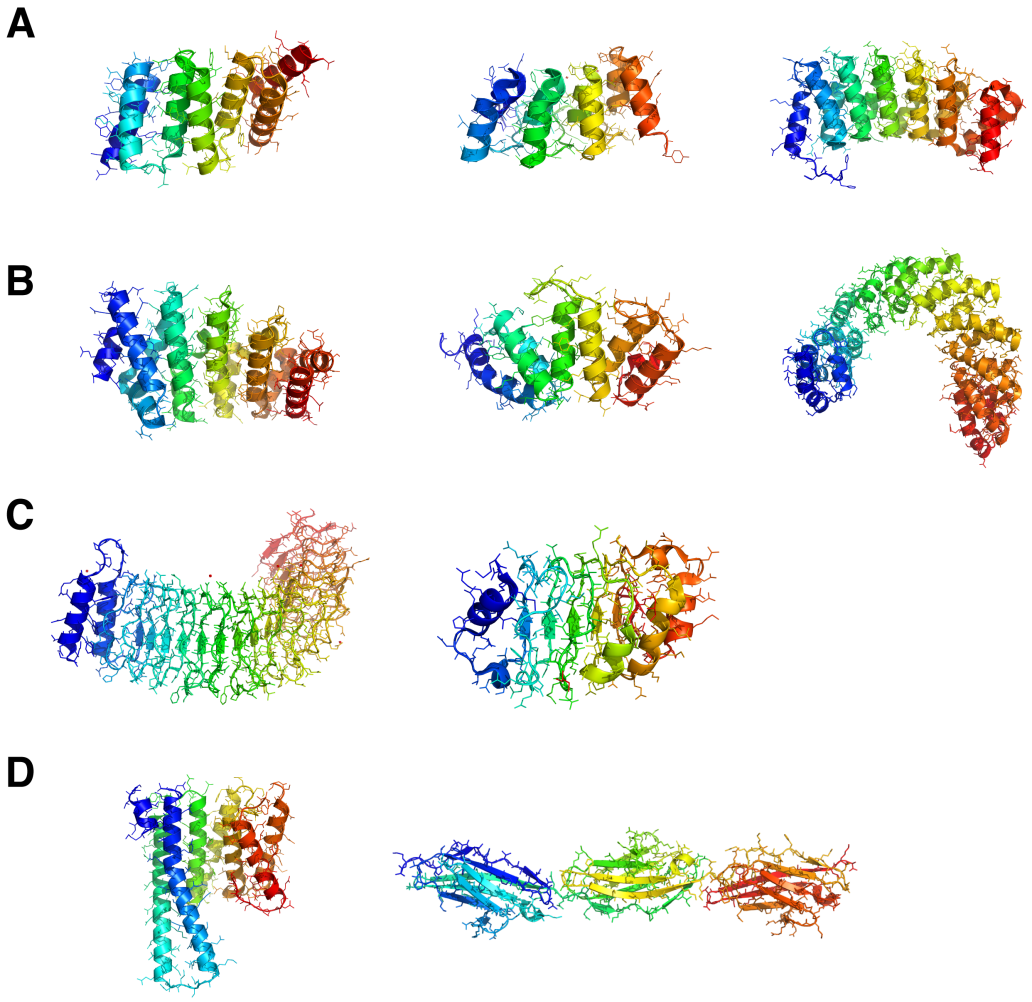


Figure 1.3. Examples of repeat and repeat-like proteins. All lists of structures are from left to right. (A) Consensus designed α-helical repeat protein motifs: consensus tetratricopeptide repeat (cTPR/c34PR; 1NA0), consensus ankyrin (cANK; 1N0R), consensus thermostable HEAT repeats (cHEAT; 3LTJ). (B) Naturally occurring α-helical repeat proteins: Armadillo repeats (4DB6), Sel1-like repeats (1KLX), TAL-effector repeats (3V6P). (C) Naturally occurring β-stranded repeat proteins: YopM LRRs (1JL5), PP32 (2JE0). (D) Proteins with repetitive architectures: 14-3-3 protein (3EFZ) and titin domains (2RIK).

In addition to topological similarity, their folding behaviors are analogous to MDPs. In recent years there have been beautiful studies on MDP folding (Batey et al., 2008; Han et al., 2007). A variety of equilibrium unfolding profiles are possible, and their features depend on the stabilities of each domain, and their interactions with each other. For example, if two isolated domains fold independently and have different stabilities, the equilibrium denaturation profile of those domains in tandem will shift and become sharper if the two domains couple and fold as a single cooperative unit (Han et al., 2007).

Repeat proteins often display this same phenomenon, as unfolding profiles of many examples display single cooperative transitions (Zweifel and Barrick, 2001; Bradley and Barrick, 2002; Courtemanche and Barrick, 2008; Dao et al., 2014). The utility of this observation is that they are also easily engineered—and can often to tolerate insertions or deletions of structural pieces[1] (Tripp and Barrick, 2004; Vieux and Barrick, 2011). This feature allows their folding to be described using one-dimensional Ising (nearest-neighbor) models to understand cooperativity in folding (Figure 1.3).

[1] The distinction must be made for “insertions” here, as this refers to the addition of thermodynamically coupled units of structure. This is in opposition to domain insertions, which may or may not be tolerated and display thermodynamic coupling. In the later case, however, repeat

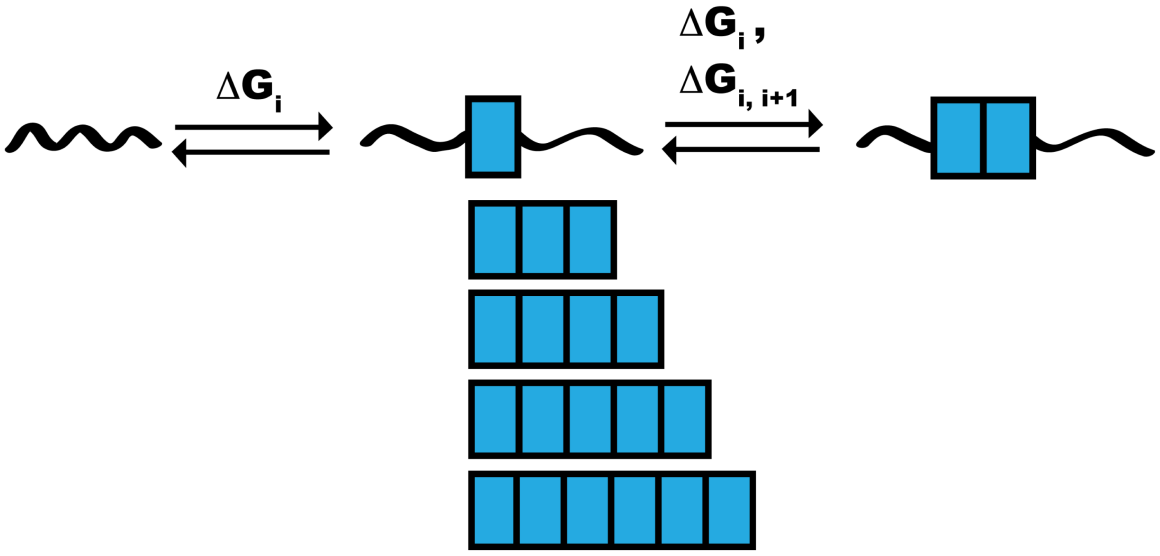


Figure 1.4. Representation of a homopolymeric Ising model for repeat protein folding. Each rectangle corresponds to a single repeat modeled as an Ising spin. Folding of single repeats is associated with an intrinsic free energy term ΔG_i, while interaction of two folded repeats yields a $\Delta G_{i,i+1}$ term. These parameters can be obtained by studying a series of proteins which vary in their numbers of repeats.

1.4 Ising analysis of repeat protein systems

The application of Ising models to repeat protein folding generally requires each repeat to have identical sequence. This criterion has been met by designing proteins based on conservation in a specific repeat family (Forrer et al., 2004; Tripp and Barrick, 2007). To date, Ising analysis has been performed on only a few repeat protein motifs, namely consensus ankyrins (cANKs) and tetratricopeptide repeats (cTPRs/c34PRs)[2] (Aksel et al., 2011; Kajander et al., 2005; Mello and Barrick, 2004; Wetzel et al., 2008). In these studies, a range of energies were found and reviewed in (Kloss et al., 2008). While cANKs are characterized by very unfavorable intrinsic repeat folding and highly stabilizing interfacial interactions, c34PRs were found to have more modest magnitudes of these parameters. We wanted to explore the molecular nature of these differences, and provide a thermodynamic description for the reduced cooperativity in c34PRs compared to ankyrins.

[2] The name TPR is derived from the prefix “tetra”, meaning four. This cannot capture variation in the tens digit, and therefore we adopt a nomenclature more intuitive of sequence length – nPR, where n corresponds to the number of residues in the repeating unit.

1.5 Overview

The structure of a 34PR single repeat consists of a pair of anti-parallel “A” and “B” helices. Since its discovery, 34PR encoding sequences have been found in a large number of genes as an interaction motif, and has been found in many proteins associated with human disease (D’Andrea, 2003).

The Pfam 27.0 database contains over 100 families classified as TPR (Finn et al., 2014). Due to similarities in their sequence motifs, TPRs have been grouped into families lacking functional annotation. This is fitting, given that repeat proteins are extremely functionally diverse (Andrade et al., 2001a, 2001b; Coates, 2003; D’Andrea, 2003; Kajava, 2001; Kobe and Kajava, 2001; Li et al., 2006; Mittl and Schneider-Brachert, 2007), yet grouping is difficult due to overlapping sequence similarity. For TPRs, this has resulted in drastically different hidden markov model (HMM) sequence representations (Figure 1.5).

In Chapter 2, I extend the nearest-neighbor modeling presented in Figure 1.4 to include additional thermodynamic terms. I use this to analyze c34PR constructs and am able to resolve single s-helix energies. In Chapter 3, I present the structure of a naturally occurring 42PR protein with a longer sequence motif, and provide evidence that it folds more cooperatively than canonical 34PRs. Finally, in Chapter 4 I apply an Ising model to quantify the cooperativity in the 42PR system from Chapter 3,

and study the effects of a key hydrogen bond in stabilizing interfacial interactions.

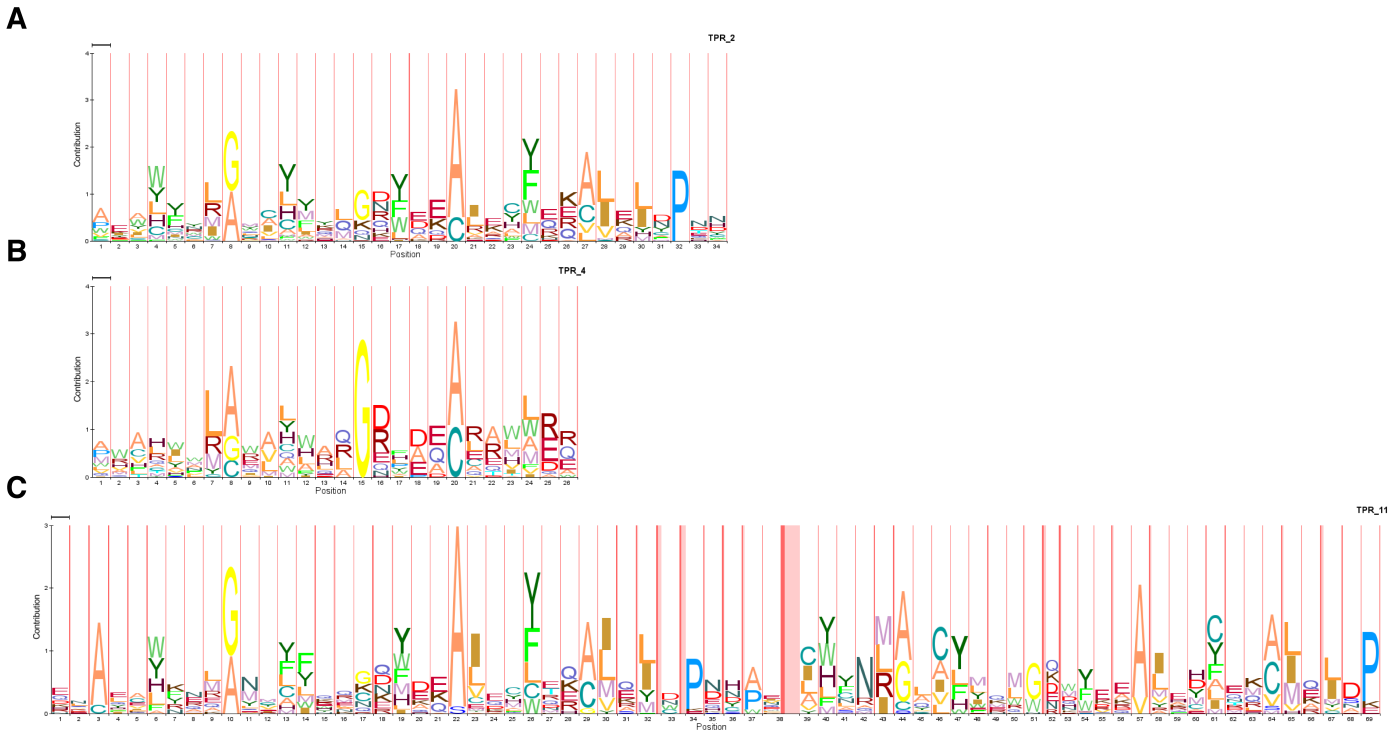


Figure 1.5. HMM logos for Pfam 27.0 classified TPR families. (A) TPR_2 34-residue logo generated from 369 seed sequences. (B) TPR_4 26-residue logo generated from 143 seed sequences. (C) TPR_11 logo generated from 458 seed sequences. This motif is 68 residues in length, which is exactly two times the canonical 34-residue motif. All logos are aligned in regions showing the greatest identity between TPR families to illustrate the frame for interpretation of the motif. Logos were obtained from Pfam 27.0 (Finn et al., 2014).

1.6 References

Aksel, T., and Barrick, D. (2009). Chapter 4 Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models. In Methods in Enzymology, (Elsevier), pp. 95–125.

Bang, M.-L., Centner, T., Fornoff, F., Geach, A.J., Gotthardt, M., McNabb, M., Witt, C.C., Labeit, D., Gregorio, C.C., Granzier, H., et al. (2001). The Complete Gene Sequence of Titin, Expression of an Unusual 700-kDa Titin Isoform, and Its Interaction With Obscurin Identify a Novel Z-Line to I-Band Linking System. Circulation Research *89*, 1065–1072.

Barrick, D. (2009). What have we learned from the studies of two-state folders, and what are the unanswered questions about two-state protein folding? Physical Biology *6*, 015001.

Cantor C.R. and Schimmel P.R. Biophysical Chermistry. WH Freeman, San Francisco, 1980.

Capp, J.A., Hagarman, A., Richardson, D.C., and Oas, T.G. (2014). The Statistical Conformation of a Highly Flexible Protein: Small-Angle X-Ray Scattering of S. aureus Protein A. Structure *22*, 1184–1195.

Chan HS, Bromberg S, Dill KA. Models of cooperativity in protein folding Phil. Trans. R. Soc. Lond. B April 1995 Volume: 348 Issue: 1323

Freire E (1995) Differential scanning calorimetry. Methods Mol Biol 40:191–218

Changeux, J.-P. (2012). Allostery and the Monod-Wyman-Changeux Model After 50 Years. Annual Review of Biophysics *41*, 103–133.

Cui, Q., and Karplus, M. (2008). Allostery and cooperativity revisited. Protein Science *17*, 1295–1307.

Cunha, E.S., Hatem, C.L., and Barrick, D. (2013). Insertion of Endocellulase Catalytic Domains into Thermostable Consensus Ankyrin Scaffolds: Effects on Stability and Cellulolytic Activity. Applied and Environmental Microbiology *79*, 6684–6696.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature 227: 561-563.

Dill, K.A. (1985). Theory for the folding and stability of globular proteins. Biochemistry *24*, 1501–1509.

Dill, Ken A. Molecular driving forces: statistical thermodynamics in chemistryphysics, biology and nanoscience. 2010 ISBN 978-0-81534430

Doig, A.J. (2002). Recent advances in helix–coil theory. Biophysical Chemistry *101*, 281–293.

Doig, A.J., Andrew, C.D., Cochran, D.A., Hughes, E., Penel, S., Sun, J.K., Stapley, B.J., Clarke, D.T., and Jones, G.R. (2001). Structure, stability and folding of the-helix. In Biochem. Soc. Symp, pp. 95–110.

Eisenberg D and Kauzmann W (1969). The Structure andProperties of Water. Oxford: Oxford University Press, NY.

Englander, S.W., Mayne, L., and Rumbley, J.N. (2002). Submolecular cooperativity produces multi-state protein unfolding and refolding. Biophysical Chemistry *101*, 57–65.

Englander, S.W., Mayne, L., and Krishna, M.M.G. (2007). Protein folding and misfolding: mechanism and principles. Quarterly Reviews of Biophysics *40*.

Fersht, A.R. (2008). From the first protein structures to our current knowledge of protein folding: delights and scepticisms. Nature Reviews Molecular Cell Biology *9*, 650–654.

Gao, M.J., Wang, J.L., Cong, Q., Zhang, B., He, X.C., Ma, X.F., and Li, G. (2015). Functionalization of Smart Gels with Beta-Cyclodextrin and Release Characteristics to Simulating Drugs. Materials Science Forum *815*, 675–683.

Gerstein M and Levitt M (1998) Simulating water and the molecules of life.Scientific American 279:100–105

Goebl M. and Yanagida M (1991). The *TPR snap helix*: a novel protein repeat motif from mitosis to transcription. Trends Biochem Sci. 16(5):173-7.

Griko, Y.V., Makhatadze, G.I., Privalov, P.L., and Hartley, R.W. (1994). Thermodynamics of barnase unfolding. Protein Science *3*, 669–676.

Grove, T.Z., Osuji, C.O., Forster, J.D., Dufresne, E.R., and Regan, L. (2010). Stimuli-Responsive Smart Gels Realized via Modular Protein Design. Journal of the American Chemical Society *132*, 14024–14026.

G. Hühne, W. Hemminger, H.J. Flammersheim. Differential scanning calorimetry. Springer, Berlin (1996)

Honig, B. (1999). Protein folding: from the levinthal paradox to structure prediction. Journal of Molecular Biology *293*, 283–293.

Jones, B.E., Jennings, P.A., Pierre, R.A., and Matthews, C.R. (1994). Development of nonpolar surfaces in the folding of Escherichia coli dihydrofolate reductase detected by 1-anilinonaphthalene-8-sulfonate binding. Biochemistry *33*, 15250–15258.

Koshland Jr, D.E., Nemethy, G., and Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits*. Biochemistry *5*, 365–385.

LiCata, V.J., and Liu, C.-C. (2011). Analysis of Free Energy Versus Temperature Curves in Protein Folding and Macromolecular Interactions. In Methods in Enzymology, (Elsevier), pp. 219–238.

Lindorff-Larsen, K., Piana, S., Dror, R.O., and Shaw, D.E. (2011). How Fast-Folding Proteins Fold. Science *334*, 517–520.

Liu Y, Sturtevant JM. Significant discrepancies between van't Hoff and calorimetric enthalpies. II. Protein Sci. 1995 Dec;4(12):2559–2561.

Liu Y, Sturtevant JM. Significant discrepancies between van't Hoff and calorimetric enthalpies. III. Biophys Chem. 1997 Feb 28;64(1-3):121–126.

Martin, R.B. (1996). Comparisons of indefinite self-association models. Chemical Reviews *96*, 3043–3064.

Matsumoto, A., and Miyahara, Y. (2014). Current Development Status and Perspectives of Self-Regulated Insulin Delivery Systems: A Review. Electronics and Communications in Japan *97*, 57–61.

Matthews, B.W. (2012). The Bragg legacy: early days in macromolecular crystallography. Acta. Cryst. Sec. A. 69, 34-36.

McKnight, J.C., Doering, D.S., Matsudaira, P.T., and Kim, P.S. (1996). A thermostable 35-residue subdomain within villin headpiece. Journal of Molecular Biology *260*, 126–134.

Mello, C.C., and Barrick, D. (2004). An experimentally determined protein folding energy landscape. Proceedings of the National Academy of Sciences of the United States of America *101*, 14102–14107.

Meng, J., Vardar, D., Wang, Y., Guo, H.-C., Head, J.F., and McKnight, C.J. (2005). High-Resolution Crystal Structures of Villin Headpiece and Mutants with Reduced F-Actin Binding Activity † , ‡. Biochemistry *44*, 11963–11973.

Moffitt, J.R., Chemla, Y.R., Smith, S.B., and Bustamante, C. (2008). Recent Advances in Optical Tweezers. Annual Review of Biochemistry *77*, 205–228.

Monod J, Wyman J, Changeux J-P. 1965. On the nature of allosteric transitions: a plausible model. J. Mol. Biol.12:88–118

Pascarella, S. and Argos, P. (1992). Analysis of Insertions/Deletions in protein structures. J. Mol. Biol. 224:461-471.

Perutz, M.F. 1970. Stereochemistry of cooperative effects in haemoglobin. Nature 228 726–739.

Prabhu, N.V., and Sharp, K.A. (2005). HEAT CAPACITY IN PROTEINS. Annual Review of Physical Chemistry *56*, 521–548.

Privalov, P.L., and Dragan, A.I. (2007). Microcalorimetry of biological macromolecules. Biophysical Chemistry *126*, 16–24.

Privalov PL. Stability of proteins: small globular proteins. Adv Protein Chem. 1979;33:167–241

Qian, H. (2012). Cooperativity in Cellular Biochemical Processes: Noise-Enhanced Sensitivity, Fluctuating Enzyme, Bistability with Nonlinear Feedback, and Other Mechanisms for Sigmoidal Responses. Annual Review of Biophysics *41*, 179–204.

Rivas, G., Fernández, J.A., and Minton, A.P. (2001). Direct observation of the enhancement of noncooperative protein self-assembly by macromolecular crowding: indefinite linear self-association of bacterial cell division protein FtsZ. Proceedings of the National Academy of Sciences *98*, 3150–3155.

Schellman, J.A. (1958). The factors affecting the stability of hydrogen-bonded polypeptide structures in solution. The Journal of Physical Chemistry *62*, 1485–1494.

Sela, M., White Jr, F.H., and Anfinsen, C.B. (1957). Reductive cleavage of disulfide bridges in ribonuclease. Science *125*, 691–692.

Sharp, K.A. (2001). Water: Structure and properties. eLS.

Sharp, K.A., and Madan, B. (1997). Hydrophobic effect, water structure, and heat capacity changes. The Journal of Physical Chemistry B *101*, 4343–4348.

Shea, M.A., and Ackers, G.K. (1985). The O R control system of bacteriophage lambda: A physical-chemical model for gene regulation. Journal of Molecular Biology *181*, 211–230.

Shortle, D. and Sondek J. (1995). The emerging role of insertions and deletions in protein engineering. Current Opinion in Biotechnology 6:387-393.

Sontag, C.., Stafford, W.., and Correia, J.. (2004). A comparison of weight average and direct boundary fitting of sedimentation velocity data for indefinite polymerizing systems. Biophysical Chemistry *108*, 215–230.

Stanley, H.E. Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, July, 1987. ISBN-10: 0195053168, ISBN-13: 9780195053166.

Vashist, S.K., Lam, E., Hrapovic, S., Male, K.B., and Luong, J.H.T. (2014). Immobilization of Antibodies and Enzymes on 3-Aminopropyltriethoxysilane-Functionalized Bioanalytical Platforms for Biosensors and Diagnostics. Chemical Reviews *114*, 11083–11130.

Wilkins, S.W. (2013). Celebrating 100 years of X-ray crystallography. Acta Cryst. Sec. A. 69, 1-4.

Zimm, B.H., and Bragg, J.K. (1959). Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. The Journal of Chemical Physics *31*, 526.

Zweifel M.E. and Barrick, D. (2001). Studies of the ankyrin repeats of the *Drosophila melanogaster* notch receptor. 2. Solution stability and cooperativity of unfolding. Biochemistry 40:14357-14367.

CHAPTER 2

Resolving stability distributions in consensus tetratricopeptide repeats (c34PRs): a heterogeneous energetic description of cooperativity in protein folding

2.1 Abstract

Many proteins fold in a highly cooperative manner, with undetectable intermediates at equilibrium. This feature of cooperativity has made it difficult to develop a complete understanding of how energy is distributed within and among elements of secondary structure. In recent years, the application of nearest-neighbor (Ising) models to repeat protein folding has enabled an understanding of cooperativity based on the interplay between two energetic terms: ΔG_i which describes the free energy change in folding a single repeat, and $\Delta G_{i:i+1}$ which describes the free energy of interaction between two adjacent folded repeats. Here, we expand this description by analyzing half-repeat units (the A and B helices), and both *inter*- and *intra*-repeat interactions. Using a system of consensus tetratricopeptide repeats (c34PRs) we resolve energetics to the

level of single α-helices (A, B, and S helices). This approach yields five (or more) uniquely determined terms (ΔG_A, ΔG_B, ΔG_S, $\Delta G_{Ai:Bi}$, and $\Delta G_{B:Ai+1}$) corresponding to different intrinsic and interfacial energies. Surprisingly, we find a rather homogeneous energy distribution at the interfacial level, despite significant differences in helix sequences and packing arrangements. At the intrinsic level, the S-helix is significantly destabilized relative to both A and B-helices. Constructs containing an S-helix show a considerable intermediate population in the absence of denaturant, which increases into the unfolding transition. Collectively, this analysis results in a more complete experimental determination of the c34PR energy landscape. Modest cooperativity in c34PRs arises from a heterogeneous distribution of moderately unstable intrinsic (ΔG_A, ΔG_B, ΔG_S) units which are offset by a homogeneous set of modest coupling energies ($\Delta G_{Ai:Bi}$, and $\Delta G_{B:Ai+1}$). The level of cooperativity observed in c34PRs is considerably lower than that observed in ankyrin, *Pa* 42PRs (see Chapter 4), and leucine rich repeat systems (Dao, 2014). These findings show cooperativity to be directly related to the magnitude of interfacial coupling and intrinsic repeat stability and provide a methodology to determine energy distributions in other similarly divisible protein systems.

2.2 Introduction

Cooperative phenomena are present at all levels of life. Macroscopically, bird and fish populations display spectacular coordinated movement (Ballerini et al., 2008; Hildenbrandt et al., 2010; Viscido et al., 2005). At microscopic levels, chemical phase transitions shift systems across narrow critical divides (Stanley, 1987). A unifying principle in these examples is a form of cooperative interaction between system components.

A major goal in modern biophysics is to thermodynamically quantify cooperativity in macromolecular systems (Baldwin, 2007). Describing cooperativity in energetic terms enables a deep understanding of complex biological processes, such as protein folding (Sosnick and Barrick, 2011), efficient hemoglobin mediated oxygen transport (Akers, 1998), and how ligand binding can determine different cellular signaling outcomes (Motlagh and Hilser, 2012).

Historically, measuring cooperative interactions in biological macromolecules has been challenging. From a theoretical perspective, the equations describing cooperativity can take simple forms, provided thermodynamic information of individual components can be directly measured. However, experimentally subdividing systems of interest into measurable pieces is extremely difficult. This is especially true in protein folding.

Recently, repeat proteins have proved to be excellent systems to understand and quantify thermodynamic origins of cooperativity in protein folding. Repeat proteins are formed from sequential arrays of modular structural units (Kajava, 2001; Kloss et al., 2008; Main et al., 2005a). The application of one dimensional nearest-neighbor (Ising) models to systems of repeat unfolding transitions (Aksel and Barrick, 2009a) has enabled descriptions of cooperativity based on the interplay of two oppositely signed energy terms, ΔG_i and $\Delta G_{i,i+1}$. These terms correspond to intrinsic repeat folding and interfacial coupling between adjacently folded repeats, respectively, and have been used to characterize consensus versions of two repeat systems – ankyrins (cANKs) (Aksel et al., 2011a; Wetzel et al., 2008) and tetratricopeptide repeats (cTPRs/c34PRs)[3] (Kajander et al., 2005a) in great detail.

One particularly interesting observation is the range of measured ΔG_i and $\Delta G_{i,i+1}$ in these systems (Kloss et al., 2008). cANKs are characterized by a very unfavorable ΔG_i offset by strong $\Delta G_{i,i+1}$, whereas c34PRs have ΔG_i near zero for whole repeats, and more modest $\Delta G_{i,i+1}$. Structurally, both systems have similarly sized repeats and bury similar amounts of solvent accessible surface area (SASA) upon folding (Kloss et al., 2008). The c34PR energies measured by Regan and coworkers are

[3] The name TPR is derived from the prefix "tetra", meaning four. This cannot capture variation in the tens digit, and therefore we adopt a nomenclature more intuitive of sequence length – nPR, where n corresponds to the number of residues in the repeating unit.

rather striking, as they suggest single repeats can fold in the absence of neighbors, a feature uncharacteristic of repeat proteins (Kajander et al., 2005).

In this study, we sought to uniquely determine the energetics of three α-helical half-repeat units (ΔG_A, ΔG_B, ΔG_S) and two interfacial interactions ($\Delta G_{Ai:Bi}$, and $\Delta G_{B:Ai+1}$) to understand their contributions to c34PR folding cooperativity. To do this, we developed an extended Ising model to resolve these energetic components.

To uniquely solve for these parameters, we created four varieties of c34PR constructs (Figure 2.1) and studied their solution properties and equilibrium unfolding. By globally fitting an extended Ising model to a large number of unfolding transitions, we find helices to be intrinsically unstable and interfaces to be stabilizing. The stabilities of the A and B-helices differ by ~0.4 kcal/mol, and the S-helix is significantly lower than A and B by ~1 kcal/mol. Surprisingly, the energies of A_i:B_i and B_i:A_{i+1} interfaces are of similar magnitude, despite significant differences in helix-helix packing and hydrogen bonding interactions. The instability of the S-helix gives rise to a substantial intermediate population at 0 M denaturant, which increases into the transition region. Interestingly, individual helices in c34PRs are on average ~4 kcals/mol more stable, despite the fact they are half the size, as cANK whole repeats (Aksel et al., 2011) or Notch Ankyrin repeats (NANKs) (Mello and Barrick, 2004). Altogether, the heteropolymeric

strategy and extended Ising model developed here provide a powerful framework to uniquely determine thermodynamic parameters which quantify cooperativity in other biomolecular systems, which need not be linear and repetitive.

2.3 Results

Design of c34PR constructs with alternating single helices

The tetetratricopeptide repeat (TPR/34PR) is a 34-residue motif present in all kingdoms of life, is composed of a pair of anti-parallel A and B-helices, and functions to mediate a variety of protein interactions (D'Andrea, 2003; Das et al., 1998; Kajava, 2001; Kloss et al., 2008; Sikorski et al., 1991). Regan and coworkers have designed a consensus version of this sequence based on a multiple sequence alignment (MSA) of over 3000 34PR sequences. Their consensus sequence was constructed by selecting the most conserved residue at each of the 34 positions in the motif (Main et al., 2003).

To increase solubility, the authors also designed a C-terminal, solubilizing S-helix by substitution of hydrophobic residues on the solvent exposed (C-terminal) face of the A-helix. Since A-helices would naturally pair next to B helices, they created constructs of the identity $(AB)_xS$, where x signifies integral numbers of whole AB units. The successful addition of single S-helices is intriguing, as it suggests c34PR helices are able to

couple to their neighbors, without the requirement for a cognate helix to make an AB (or SB in this example) unit. Therefore, we sought to see if this same strategy could be implemented using native (A,B) helices on both the N- and C-terminal ends of c34PR arrays, which would allow us to resolve the contributions of each helix to stability.

To create c34PR constructs which vary in the length and ratio of A, B, and S helices, we used the Regan consensus 34PR sequence as a guide to generate individual DNA cassettes encoding for AB, A, B, and S sequences (Figure 2.1A). We designed DNA cassettes to have flanking, complementary, BamHI and BglII "sticky-end" restriction sites to allow for construct elongation. This strategy has been used in other studies of repeat proteins (Aksel et al., 2011a; Javadi and Main, 2009; Carrion-Vazquez et al., 1999; Hongbin and Fernandez, 2003).

For each integral number of central AB repeat, there are six possible construct architectures: $(AB)_x$, $B(AB)_x$, $(AB)_xA$, $(AB)_xS$, $B(AB)_xA$, and $B(AB)_xS$, where x represents integer increments of whole AB units. Due to self-association, only a subset of these construct architectures were used in this study (Figure 2.1B). In total, we constructed 15 different c34PR proteins from these architectures, which range in length from one to four central AB units.

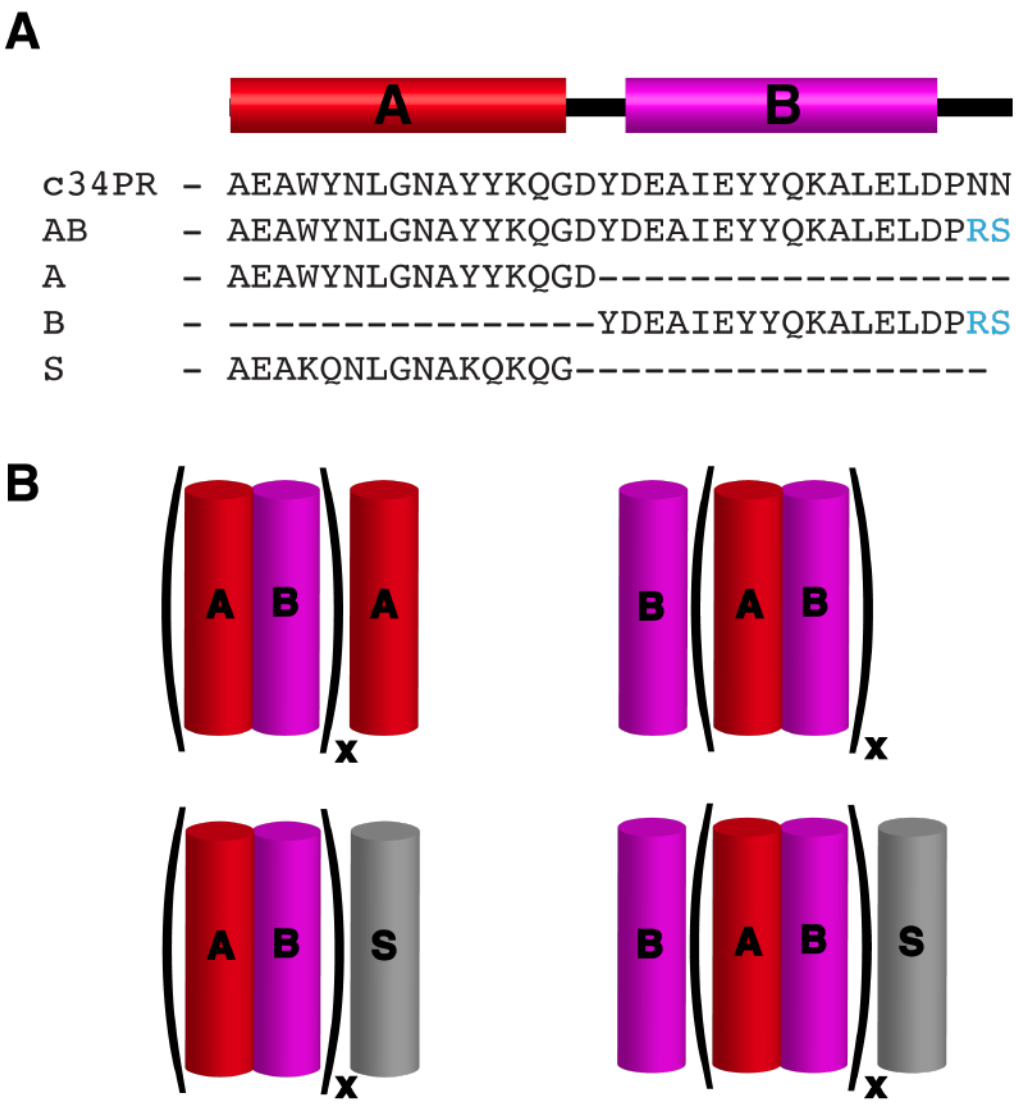


Figure 2.1. c34PR sequences and construct architectures used in study. (A) Sequences corresponding to A, B, and S helices are aligned according to their position in the c34PR consensus sequence. RS substitutions for cloning purposes are colored in blue. Helix boundaries are from the structure 1NA0. (B) Construct architectures used in this study, where x represents the number of internal AB repeats, ranging from one to four.

CD spectroscopy of c34PR constructs

The c34PR constructs studied by Regan and coworkers displayed α-helical CD spectra consistent with their structures (Main et al., 2003). Therefore, we expected our $B(AB)_xS$ and $B(AB)_x$ constructs to display α-helical far-UV spectra. We find all c34PR constructs to have α-helical far-UV CD spectra characteristic of canonical c34PR folds (Figure 2.2). Decreased molar residue ellipticities (MRE) are observed for the shortest constructs ABS and BAB. This decrease could potentially reflect the instability of these constructs compared to the rest of the series. The remaining constructs have similar spectral shapes. Slight deviations in MRE values for the longer proteins are likely to be due to concentration uncertainty. The MRE magnitudes observed here are consistent with values previously reported for c34PRs which lack single helical additions beyond C-terminal S-helices (Main et al., 2003).

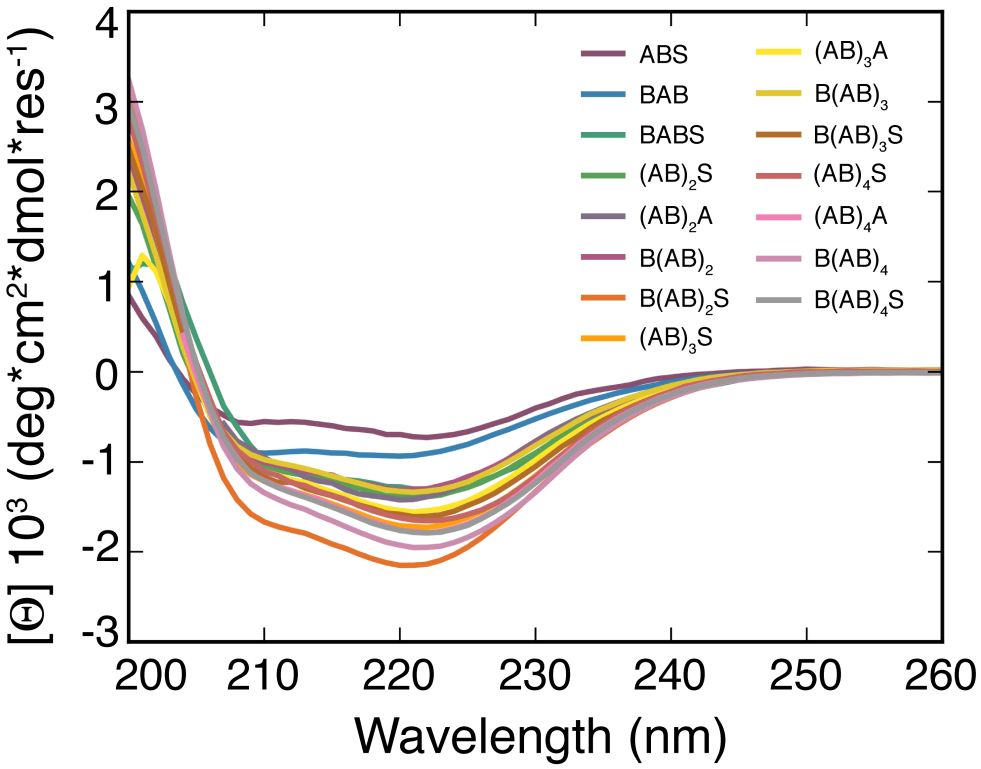


Figure 2.2. Far-UV spectra of c34PR constructs, collected in 50mM Na Phosphate, 150mM NaCl, pH 6.8 at 25°C. Constructs in legend are displayed in order of increasing length and are characterized by a minimum at 220 nm.

Solution NMR spectroscopy of c34PRs

To see if our c34PR constructs have well-defined tertiary structures, we collected ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) nuclear magnetic resonance (NMR) spectra on a representative set of proteins (Figure 2.3). While most peaks in the spectra are well defined, some show moderate broadening. In addition, there are fewer peaks in each construct than would be expected from the number of non-proline residues in each protein. A possible explanation for this behavior is self-association.

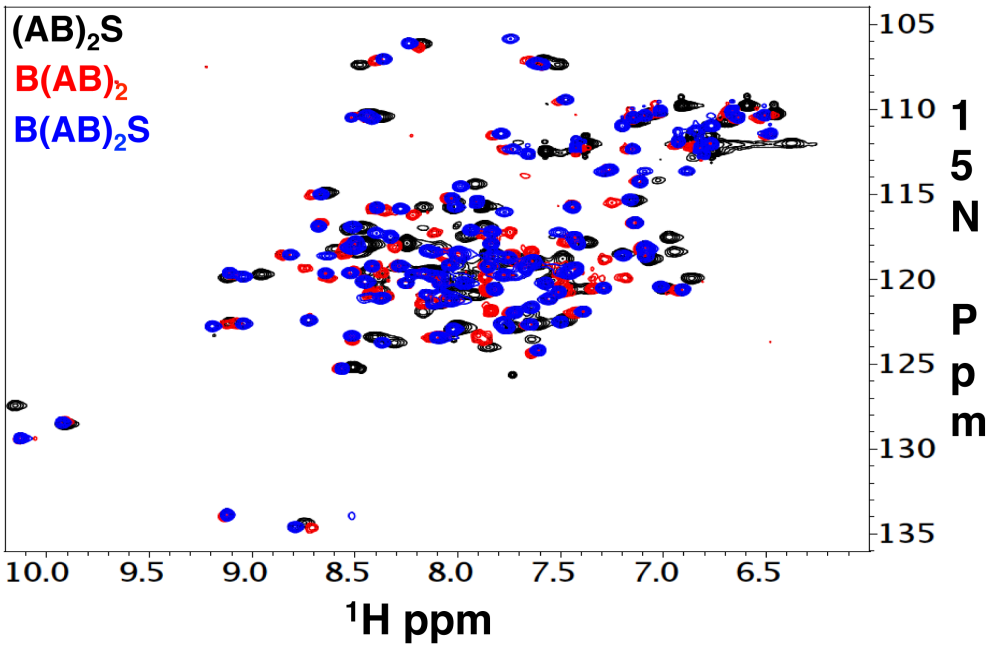


Figure 2.3. ^{1}H-^{15}N HSQC NMR spectra of representative c34PR constructs. Spectral colors for each construct are indicated in the legend. Experimental conditions and data collection strategy are outlined in the supplementary material.

Analytical ultracentrifugation sedimentation velocity of c34PRs

To determine if our designed c34PR constructs self-associate in solution we performed sedimentation velocity analytical ultracentrifugation (SV-AUC) experiments. We analyzed SV data using direct boundary ($\Delta C/\Delta T$) methods (Stafford and Sherwood, 2004), as well as c(s) methods (Schuck, 2000) (data not shown). Although previous sedimentation equlibrium studies have found c34PRs to be monomeric (Main et al., 2003), we find constructs to weakly self-associate. These differences may result from the low sensitivity of sedimentation equilibrium methods to small amounts of aggregates or the cloning substitutions between repeats (Figure 2.1). Other kinetic and equilibrium studies of c34PRs containing these substitutions did not mention potential consequences of these substitutions (Javadi and Main, 2009). The construct architectures displayed in Figure 2.1B are predominantly monomeric in solution.

Although the $(AB)_x$ and $B(AB)_xA$ constructs express well and are soluble, they form large oligomers in solution as analyzed by SV-AUC. Associations persist even at low (<5μM) concentrations and in the presence of native baseline level concentrations of urea (not shown). It is interesting to note that, despite these associations, the unfolding transitions for these constructs display single, cooperative, transitions with

m-values and stabilities that are appropriate in magnitude for constructs of their size and helical composition at low concentrations. When concentrations are increased, unfolding transitions shift in stability, yet retain similar m-values to transitions collected at low concentrations. The most consistent explanation for this association behavior seems to be due to structurally compatible N and C-terminal helices. In these constructs, the solvent exposed N-terminal helix faces are poised to pair with cognate C-terminal helix faces from another molecule, and vise-versa.

From a theoretical and physical perspective, it is also interesting to think about the nature of the equilibrium constants involved in the potential end-to-end associations. For example, is the addition of each monomer characterized by the same equilibrium constant (isodesmic)? Perhaps the Enthalpic contribution of each monomer is equal (Chatelier, 1987)? Indefinite association schemes such as these have been found for some systems (Sontag et al., 2004), and have some have even been implemented as fitting models for the analytical ultracentrifuge. Despite these interesting considerations, based on our inability to determine the true oligomeric state of these constructs, we have excluded them from further analysis.

Thermodynamic stability of c34PRs

To measure the thermodynamic stability of our designed c34PR constructs, we conducted urea-induced equilibrium unfolding titration unfolding. All c34PR constructs display single, cooperative unfolding transitions, which are completely reversible. As helices are added to the arrays, the unfolding midpoints shift to higher denaturant conentrations. Moreover, stabilities depend on the type of helices in each array (Figures S2.1 and S2.2). The constructs ABS and BAB are the least stable, and lack well-defined native baselines, consistent with their reduced MRE magnitudes (Figure 2.2). The dependence of midpoint on helix type demonstrates that the stabilities of either helices or interfaces differ among A, B, and S.

The c34PR unfolding transitions also become sharper as length is increased. For constructs containing greater than six helices, the m-values (as analyzed by a two-state unfolding model plateau around 1.5 kcal*mol^{-1}*M^{-1}. This suggests an upper limit to the size of the cooperative unit, and may be due to the presence of equilibrium intermediates.

Heteropolymeric nearest-neighbor model development

To quantify potential equilibrium intermediates in unfolding transitions, and to resolve potential differences among intrinsic and interfacial energies in c34PR arrays, we globally analyzed equilibrium

unfolding transitions with eight nearest neighbor models (M1-M8, Table 2.1). To fit our c34PR data, the traditional one-dimensional nearest-neighbor models used to analyze other repeat protein systems (Aksel and Barrick, 2009b; Aksel et al., 2011; Kajander et al., 2005b; Mello and Barrick, 2004b; Wetzel et al., 2008) must be extended to include new interface statistical weights. We developed a modeling approach to include multiple intrinsic (ΔG_A, ΔG_B, ΔG_S), and interfacial ($\Delta G_{Ai:Bi}$ and $\Delta G_{B:Ai+1}$) energy terms. These energies correspond to unique α-helix, and helix-helix packing interactions, respectively. Figure 2.4 illustrates our approach from an energetic perspective, from an $(AB)_3$ starting point.

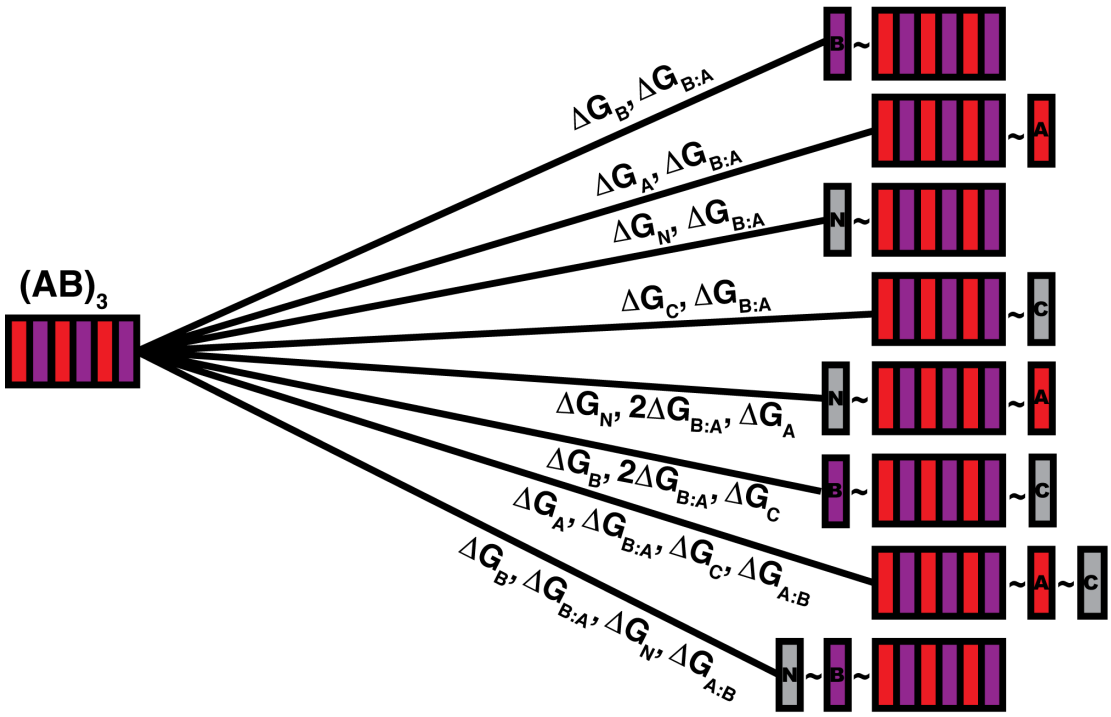


Figure 2.4. Heteropolymeric Ising approach as applied to c34PRs. This approach extends the formalism shown in Figure 1.3. Individual helices are treated as Ising spins. Constructs on the right show the addition of one or two terminal helices to an $(AB)_3$ core. Associated energy terms for each addition are shown along the connecting lines. The red, purple, and grey bars represent A, B, and different types of solvation capping helices, respectively. In c34PRs, the only capping helix used corresponds to a C-terminal variant of the native A-helix (S-helix). Studying these constructs as a function of the number of $(AB)_x$ central units provides a unique solution of the energy terms illustrated in the diagram.

To analyze unfolding transitions of c34PRs, we generated partition functions for each construct. Casting helical (K_i and K_j) and interface ($\mathrm{W}_{i:j}$ and $\mathrm{W}_{j:i}$) statistical weights in two-by-two matrices results in the following representation of a partition function (q) for a protein containing two different helices:

$$q = \begin{bmatrix} 0 & 1 \end{bmatrix} \times \left(\begin{bmatrix} \mathrm{K}_i \mathrm{W}_{j:i} & 1 \\ \mathrm{K}_i & 1 \end{bmatrix} \times \begin{bmatrix} \mathrm{K}_j \mathrm{W}_{i:j} & 1 \\ \mathrm{K}_j & 1 \end{bmatrix} \right)^{\mathrm{n}} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (2.5)$$

Here, the K and W statistical weight terms define different helices and interfaces, respectively, which are designated by their subscripts. The superscript n, corresponds to the number of consecutively paired helical matrices, as in c34PRs, an alternating pattern of A and B-helices is observed and therefore must be incorporated into the matrix multiplication scheme. For N and C-terminal “capping” helices (Figure 2.4), one needs only to include them as terminal two-by-two matrices. Only one unique capping parameter can be solved (intrinsic or interface). Often in repeat protein design, capping motifs include polar substitutions on the solvent exposed side, leaving the interface residues unchanged. Therefore, it seems a reasonable approximation to fit unique intrinsic energies of capping motifs. Including these terminal capping matrices results in the following partition function for a protein composed of n AB helix pairs, capped at both the N- and C-terminus:

$$q = \begin{bmatrix} 0 & 1 \end{bmatrix} \times \begin{bmatrix} K_N W_{i:j} & 1 \\ K_N & 1 \end{bmatrix} \times \left(\begin{bmatrix} K_i W_{j:i} & 1 \\ K_i & 1 \end{bmatrix} \times \begin{bmatrix} K_j W_{i:j} & 1 \\ K_j & 1 \end{bmatrix} \right)^n \times \begin{bmatrix} K_C W_{i:j} & 1 \\ K_C & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

(2.6)

To quantify cooperativity and potential intermediate populations in c34PRs, we constructed eight nearest neighbor models (M1-M8, Table 2.1) based on Ising theory (Ising, 1925; Poland and Scheraga, 1970). We used these models in separate global fits of the same c34PR equilibrium unfolding data. Models M1-M8 contain different combinations of shared parameters corresponding to different free energies of helices (ΔG_A, ΔG_B, ΔG_S), and interfaces ($\Delta G_{Ai:Bi}$, $\Delta G_{B:Ai+1}$) in c34PR arrays. We further increased the possible parameter combinations by including models which contain one (either m_i or $m_{i,i+1}$) or two (both m_i and $m_{i,i+1}$) denaturant sensitivities (m-values) which are ascribed to either intrinsic or interfacial energies, denoted by their subscripts. Details regarding the fitting procedure and analysis are outlined in the experimental methods section.

The different types of models are displayed in Table 2.1 (characterized by their fitted parameters), along with their fit statistics, and are ranked from lowest to highest reduced χ^2 (χ^2/ν), where ν represents the number of degrees of freedom in each fit. The models that fit the best based on reduced χ^2 (M1-M4) all contain separate fitted intrinsic energies for each helix (ΔG_A, ΔG_B, ΔG_S). Although models M4-M8 (which have a

single intrinsic free energy parameter) are able to fit the data with reasonable statistics, the fits are all considerably worse than the fits for M1-M4 (two fold higher χ^2/ν values, Table 2.1).

Table 2.1. c34PR nearest neighbor model parameters and global fit statistics

Model	#params	$\chi 2/\nu$	ΔG_A[a]	ΔG_B[a]	ΔG_S[a]	$\Delta G_{Ai:Bi}$[a]	ΔG_{BiAi+1}[a]	m_i[b]	$m_{i,i+1}$[b]
M1	6	$1.56E^{-4}$	1.60	1.18	2.33	-2.72	$\Delta G_{Ai:Bi}$	-0.54	0.39
			{1.41,1.8}[d]	{0.9,1.38}[d]	{2.14,2.52}[d]	{-2.96,-2.48}[d]		{-0.58,-0.5}[d]	{0.33,0.44}[d]
M2	7	$1.57E^{-4}$	1.60	1.18	2.33	-2.72	-2.72	-0.54	0.39
			{1.42,1.8}[d]	{0.98,1.38}[d]	{2.15,2.53}[d]	{-2.98,-2.48}[d]	{-2.97,-2.47}[d]	{-0.58,0.5}[d]	{0.33,0.44}[d]
M3	6	$1.79E^{-4}$	2.97	2.6	3.58	-4.34	-4.48	-0.224	
			(2.6,3.36)[c]	(2.25,3.0)[c]	(3.17,4.15)[c]	(5.02,-3.81)[c]	(-5.16,-3.93)[c]	(-0.24,-0.21)[c]	0
			{2.78,3.72}[d]	{2.58,3.55}[d]	{3.24,4.22}[d]	{-5.59,-3.88}[d]	{-5.83,-4.45}[d]	{-0.27,-0.21}[d]	
M4	5	$1.79E^{-4}$	2.98	2.6	3.6	-4.44		-0.22	0
M5	5	$2.7E^{-4}$	1.95	ΔG_A	ΔG_A	-3.73	-2.93	-0.44	-0.27
M6	4	$2.82E^{-4}$	2.89	ΔG_A	ΔG_A	-4.86	-4.12	-0.22	0
M7	4	$2.94E^{-4}$	2.00	ΔG_A	ΔG_A	3.38	$\Delta G_{Ai:Bi}$	-0.41	0.24 {-
			{-0.92,4.54}[d]					{-0.93,0.16}[d]	0.42,0.81}
M8	3	$3.06E^{-4}$	2.87	ΔG_A	ΔG_A	-4.37	$\Delta G_{Ai:Bi}$	0.21	
			(2.45,3.08)[c]			(-5,-3.91)[c]		(0.19-0.24)[c]	0
			{2.43,3.73}[d]			{-5.52,-3.97}[d]		{-0.26,-0.21}[d]	

#params indicates the number of relevant thermodynamic parameters in the model.

[a] Units are in kcal*mol^{-1}

[b] Units are in kcal*mol^{-1}*M^{-1}

[c] 95% F-statistics confidence intervals were calculated from analysis using equation 35 from (Johnson and Straume, 1994)

[d] 95% Bootstrap confidence intervals were calculated from 1800 iterations using the method presented in (Efron and Tibshirani, 1993) and summarized in (Johnson, 2008).

To assess the statistical significance of the χ^2/ν reduction, we computed F-statistics for each pair of models. The F-statistic is a ratio of two χ^2/ν from different models, and can be used in a probability distribution function to obtain confidence levels. Generally, a model with a greater number of parameters will result in a lower χ^2/ν compared to models with fewer parameters. As with parameter error estimation, the percentage value obtained from an F distribution corresponds to the integral under the probability distribution function, and represents the level of confidence in the increase in fit performance using one model over another.

We therefore performed F-statistics calculations for all possible pairwise model comparisons (Table 2.2). The percent confidence is displayed in parenthesis next to the F-value for each comparison. By convention, F-statistics are always greater than one. If two models have the exact same χ^2/ν, the F is 1.0. As F increases, the statistical significance of the better-fit model (denominator) over the weaker model increases.

Table 2.2. c34PR nearest neighbor model F-statistic comparison

Model	M1	M2	M3	M4	M5	M6	M7
M1							
M2	1.001 (50.6)						
M3	1.15 (98.7)	1.15 (98.7)					
M4	1.15 (98.8)	1.15 (98.7)	1.001 (50.6)				
M5	1.73[a]	1.73[a]	1.51[a]	1.51[a]			
M6	1.80[a]	1.8[a]	1.57[a]	1.57[a]	1.04 (75.7)		
M7	1.89[a]	1.89[a]	1.65[a]	1.65[a]	1.09 (92.5)	1.05 (77.2)	
M8	1.95[a]	1.7[a]	1.7[a]	1.7[a]	1.13 (97.6)	1.08 (90)	1.04 (70.4)

F-values were calculated as row/column models from a ratio of reduced $\chi 2$ values, and values in parenthesis indicate the % confidence in the better-fit model (column) from each comparison. Due to similarity in degrees of freedom between the models, the critical F values for 1σ, 2σ, and 3σ correspond to ~1.029, ~1.109, and ~1.185, respectively, for all model comparisons.

[a] An F > 1.453 corresponds to a p value of 99.$\bar{9}$%.

The F statistics in Table 2.2 show that treating helices as separate results in a significant statistical improvement corresponding to a confidence of nearly, 100%, over models where a single helix energy is used. Overall, the inclusion of separate intrinsic parameters significantly improves fit quality (compare M1-M4 with M5-M8), and including separate interface parameters does not improve fit quality (M1 vs. M2 and M3 vs. M4).

For models with separate helix terms, some interesting trends are observed. Models M3 and M4 differ by the inclusion of an additional interfacial energy term, and model the denaturant sensitivity as affecting only the intrinsic units. These two models are statistically similar to one another, as the added degree of freedom in going from M3 to M4 cannot explain the improvement in fit quality (confidence 50%).

Similarly, models M1 and M2 differ by the inclusion of an additional coupling energy term. For M1-M4, the inclusion of $m_{i,i+1}$ significantly improves the fit over models where it is absent (confidence ~98.7%). Although M1 has the lowest χ^2/ν of all models, it cannot be distinguished from M2. Still, we prefer M1, due to the lower number of parameters, and the fact that the added interfacial parameter in M2 has the same best-fit value as the one in M1. The interpretations that follow do not change based on choosing M1 over M2 (Table 2.2). The fit corresponding to M1 is

shown in Figure 2.7, and fits corresponding to models M1-M8, along with associated parameter error distributions for M1, M2, and M8, are presented in the supplementary material to this chapter (Figures S2.4-S2.5).

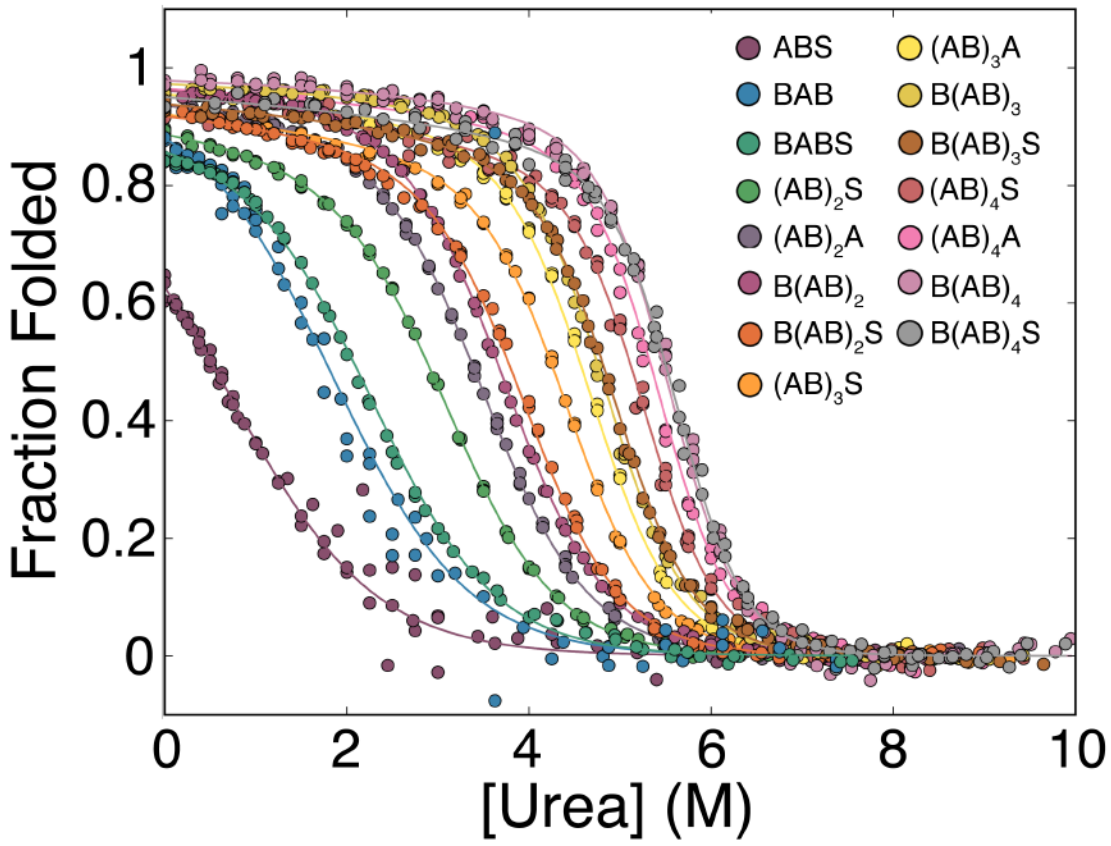


Figure 2.7. Urea-induced equilibrium unfolding and refolding transitions of c34PR constructs. Data for constructs are displayed as solid circles, and are listed in order of increasing unfolding midpoint in the legend. At least three independent titrations were performed for each construct. The solid lines result from the best-fit parameters in a global analysis method using model M1 (Table 2.1), and are colored the same as the construct data in Figure 2.2.

The parameters obtained from the fit of M1 to the data show cooperativity in c34PRs to arise through unfavorable helix energies which are offset by favorable helical coupling energies. Moreover, while all of the helices are unstable, they have different stabilities (in terms of free energy, $B < A < S$, Table 2.1).

2.4 Discussion

Traditional Ising models applied to repeat protein folding (Aksel et al., 2011a; Kajander et al., 2005a; Mello and Barrick, 2004a; Wetzel et al., 2008) have involved descriptions of systems using two terms: ΔG_i and $\Delta G_{i,i+1}$. While some of these models have the potential to resolve multiple ΔG_i terms, none of them are able to capture multiple $\Delta G_{i,i+1}$ terms. This prevents an understanding of how intrinsic units are energetically subdivided. In this study we have extended the traditional Ising modeling approach of repeat systems to include intra-repeat coupling energies.

Our motivation for creating c34PRs with different ratios of A, B, and S helices was to resolve energetics of individual helices, and intra-repeat helical packing. c34PRs have unique repeat architectures and can be well-approximated as a series of alternating α-helices. There is little sequence identity between A and B-helices, and S-helices have five polar substitutions from hydrophobic residues of the A-helix from which it was derived (Figure 2.1). In addition, the structural arrangement and packing of intra (A_i:B_i) and inter-repeat (B_i:A_{i+1}) interfaces is quite different (Figure S2.7). Therefore, it is plausible these sequence and structural differences could give rise to different energetic parameter values, and a non-uniform stability variation along the molecule.

Stability distribution in c34PRs

We find a heterogeneous distribution of stability for intrinsic units, consistent with their sequence differences. We find the B-helix to be ~0.4 and ~1 kcal/mol more stable than the A- and S-helices, respectively. Despite the packing differences between A_i:B_i and B_i:$A_{i,i+1}$ (Figure S2.7), we find both interface energies to be -2.7 kcal/mol, and our data are equally well-described when using a single interface parameter.

The magnitude of the intrinsic S-helix energy is almost equal to the magnitude of the interface coupling energy. This likely results in partial unfolding of the helix, and is consistent with previously reported hydrogen exchange NMR (HX-NMR) experiments where the S-helix was found to show no exchange protection (Main et al., 2005). In addition, the normalized CD titrations of data collected here show decreased stability for constructs containing S-helices, relative to constructs of equivalent helical number lacking the S-helix (Figure S2.1).

A summation of ΔG_A, ΔG_B and $\Delta G_{Ai:Bi}$ parameters gives the total energy of a single 34-residue consensus repeat of 0.06 kcal/mol. Therefore, the equilibrium constant for intrinsic whole repeat folding is about 1.0, and a single c34PR is equally likely to unfold as it is to fold in the absence of its neighbors. This is consistent with the observed low apparent cooperativity in c34PRs when fitting using two state models (Chapter 3).

The Ising parameters obtained here paint a much different picture of c34PR folding when compared to more cooperative systems such as cANKs, NANKs, or cLRRs, which are all characterized by much more (> 6 kcal/mol) unstable intrinsic repeats, and stronger (> 8 kcal/mol) interfacial interactions. Although they are roughly half the size of ankyrin repeats, c34PR helices are about 4 kcal/mol more stable than cANKs. Collectively, a comparison of energetics explains the much lower folding cooperativity in c34PRs versus these other systems.

The energies here are different from, although not inconsistent with, those previously reported by Regan and coworkers (Kajander et al., 2005), with a derived single c34PR intrinsic energy of 0.5 kcal/mol. In their modeling, they globally fit constructs ranging in integral numbers from 2.5 to 10.5 repeats (5-21 total helices). To model these data, they assumed equal energy of c34PR helices and interfaces. Modeling the data in this way was required because these constructs all contained an invariant S-helix, and equal numbers of A and B-helices. Their data were reasonably well described by the model; however, it lacked the ability to determine intrinsic and interfacial energy differences in c34PRs. While we find their assumption of equal energy to be valid at the interfacial level (Table 2.1) (confirmed using our data), our data show the intrinsic energies of c34PR helices to be very different. For our data set, where the ratio of different helix types differs among constructs, modeling using single helix and

interface terms (model M8) fits the poorest out of all models tested. This demonstrates that there is variation among the stabilities of the helices and interfaces in c34PRs.

Error analysis in c34PRs and other Ising modeled systems

Of importance whenever fitting models to data is an estimation of the errors associated with the fitted parameters. The approach we present here is to analyze the error distributions using F-statistic (Johnson and Straume, 1994) and bootstrapping methods (Johnson, 2008). These statistical methods are especially important for reporting errors for non-linear systems, as errors derived from the covariance matrix tend to be underestimations. In addition, the errors can be asymmetric, and covariance matrix error estimations assume normal distributions.

For models M1, M2, and M8, bootstrapped obtained error distributions are shown in Figure S2.5. The errors for models M1 and M2 are reasonably normally distributed. For models containing a single helix free energy, the error distribution is bimodal (magenta histrograms, Figure S2.5). A likely explanation for this is the failure of these models to capture the instability of the S-helix compared to the A and B helices. Since roughly half of the constructs contain an S-helix, there is a minimum for models M5-M8 to minimize mostly with respect to these in the fit, whereas another minimum exists for proper modeling of the constructs which do not

contain an S-helix. The covariance matrix errors are displayed in Table 2.3.

Ising models have proven to be very useful in determining origins of cooperativity in equilibrium folding. However, Ising models applied to repeat protein folding have a negative correlation between intrinsic and interfacial energy. These correlations are diminished when a collection of constructs can be analyzed that have large variations in construct length. Likewise, correlation between single helix terms are diminished by including constructs that vary the ratio of the helix types. An illustration of these correlations can be viewed in two-dimensional F-statistics calculated confidence regions (not shown).

It is unfortunate we cannot trust the oligomerization state of the (AB)x and B(AB)xA architectures. Using them in the global modeling would help constrain parameter values, as there would be more ways to energetically differentiate constructs from one another. In addition, the added degrees of freedom in each fit would reduce the critical F values corresponding to the confidence limits of interest, which would likely lead to tighter limits.

Table 2.3. c34PR Ising thermodynamic covariance matrix error estimations

Model	$\chi 2$	$\chi 2/\nu$	ΔG_A [a]	ΔG_B [a]	ΔG_S [a]	$\Delta G_{Ai:Bi}$ [a]	ΔG_{BiAi+1} [a]	m_i [b]	$m_{i,i+1}$ [b]
M1	0.168	$1.56E^{-4}$	1.60 ± 0.11	1.18 ± 0.11	2.33 ± 0.1	-2.72 ± 0.13	$\Delta G_{Ai:Bi}$	-0.543 ± 0.02	0.386 ± 0.03
M2	0.168	$1.57E^{-4}$	1.60 ± 0.11	1.18 ± 0.11	2.33 ± 0.1	-2.72 ± 0.13	-2.72 ± 0.14	-0.543 ± 0.02	0.386 ± 0.03
M3	0.193	$1.79E^{-4}$	2.97 ± 0.03	2.6 ± 0.04	3.58 ± 0.04	-4.34 ± 0.06	-4.48 ± 0.06	-0.224 ± 0.002	0
M4	0.193	$1.79E^{-4}$	2.98 ± 0.03	2.6 ± 0.04	3.6 ± 0.04	N/A	-4.44 ± 0.04	-0.224 ± 0.002	0
M5	0.291	$2.7E^{-4}$	1.95 ± 0.14	ΔG_A	ΔG_A	-3.73 ± 0.17	-2.93 ± 0.18	-0.436 ± 0.03	0.265 ± 0.04
M6	0.304	$2.82E^{-4}$	2.89 ± 0.04	ΔG_A	ΔG_A	-4.86 ± 0.07	-4.12 ± 0.06	-0.217 ± 0.002	0
M7	0.317	$2.94E^{-4}$	2.0 ± 0.01	ΔG_A	ΔG_A	-3.38 ± 0.18	$\Delta G_{Ai:Bi}$	-0.413 ± 0.03	0.24 ± 0.04
M8	0.329	$3.06E^{-4}$	2.87 ± 0.04	ΔG_A	ΔG_A	-4.44 ± 0.05	$\Delta G_{Ai:Bi}$	0.214 ± 0.002	0

[a] Units are in $kcal*mol^{-1}$

[b] Units are in $kcal*mol^{-1}*M^{-1}$

Errors displayed represent estimations obtained from the covariance matrix corresponding to each fit.

The energy landscape of c34PRs and intermediate populations

Using the best-fit energy values from M1, we can visualize the entire protein folding energy landscape of any c34PR array of helices. Figure 2.8 displays the experimentally determined energy landscape of $B(AB)_2S$. The native state is close in energy to a microstate lacking a folded S-helix.

A powerful feature of Ising analysis is that it allows the direct calculation of intermediate populations along folding trajectories. For c34PR constructs containing an S-helix, a substantial intermediate population is observed. This is reflected in the energy landscape in Figure 2.8. Calculation of all intermediate populations leads to a detailed picture of the equilibrium species. This can be represented in a four-dimensional plot, with coloring corresponding to the free energy of each intermediate, relative to the unfolded state, in water (Figure 2.9).

Another useful representation is to total all of the intermediates, which collapses the y-axis in Figure 2.9, summing the populations (z-axis) along the way (Figure 2.10, dashed lines). This plot shows c34PRs to have a considerable intermediate population, even at 0 M denaturant. Collectively, the heteropolymeric Ising model we applied leads to a detailed description of energy landscape of and folding cooperativity of c34PRs. The intermediate populations are significantly higher than cANKs

or cLRRs due to the increased stability of intrinsic c34PR units, and lower coupling energies between repeats. These high levels of intermediates are consistent with a mismatch between the calorimetric and van't Hoff enthalpies previously observed for c34PRs (Cortajarena and Regan, 2011), suggesting multistate equilibrium unfolding (although we were unable to reproduce thermal reversibility when trying to replicate these results).

While our modeling approach was specific to c34PRs, a similar methodology can be applied to global analyses of other systems. Thermodynamically, anytime a system can be broken into its constituent parts and studied in the absence of neighbors, parameters can be obtained that quantify the energies of interaction between units.

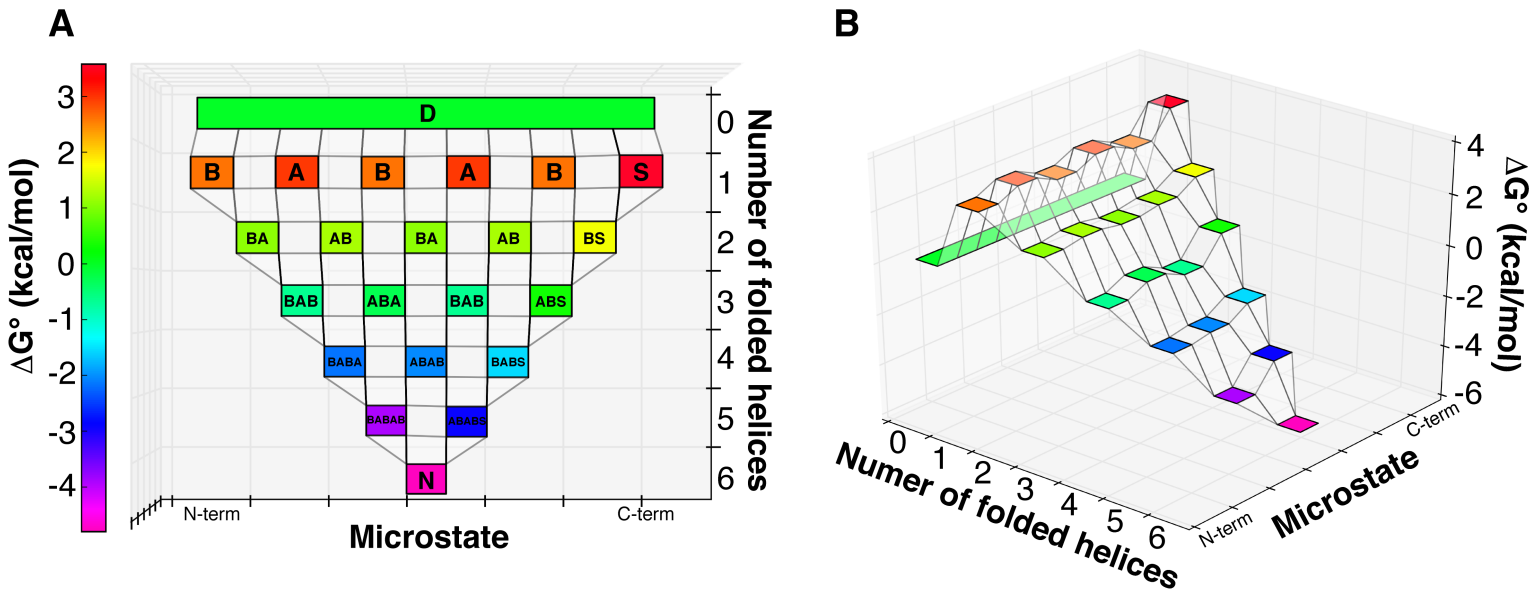


Figure 2.8. Energy landscape of $B(AB)_2S$. (A) Top view of energy landscape with microstate labels. Helices are labeled as A, B, and S. Only microstates with consecutively folded helices are displayed for clarity (B) Side view of $B(AB)_2S$ energy landscape. Microstates are colored the same as in (A).

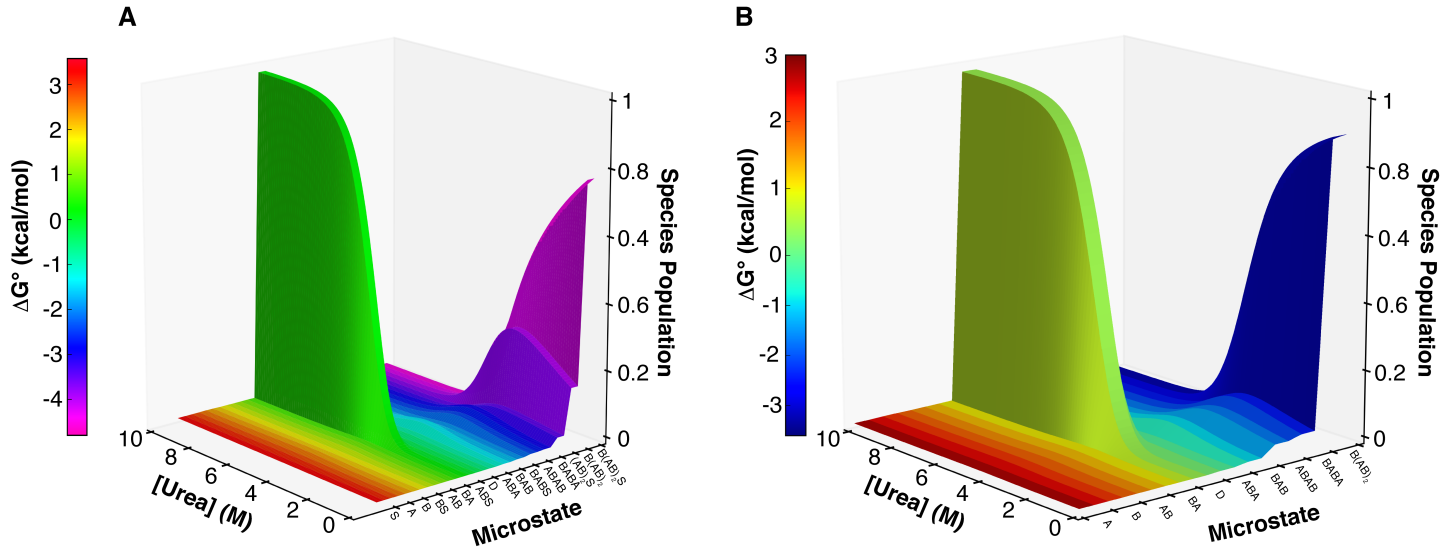


Figure 2.9. Four dimensional representation of $B(AB)_2S$ (A) and $B(AB)_2$ (B) species populations. Microstates are ordered with respect to increasing free energy and are colored to reflect free energies in the absence of denaturant as in Figure 2.8. (B) $B(AB)_2$ species population plot. Microstates are ordered and colored with respect to free energy.

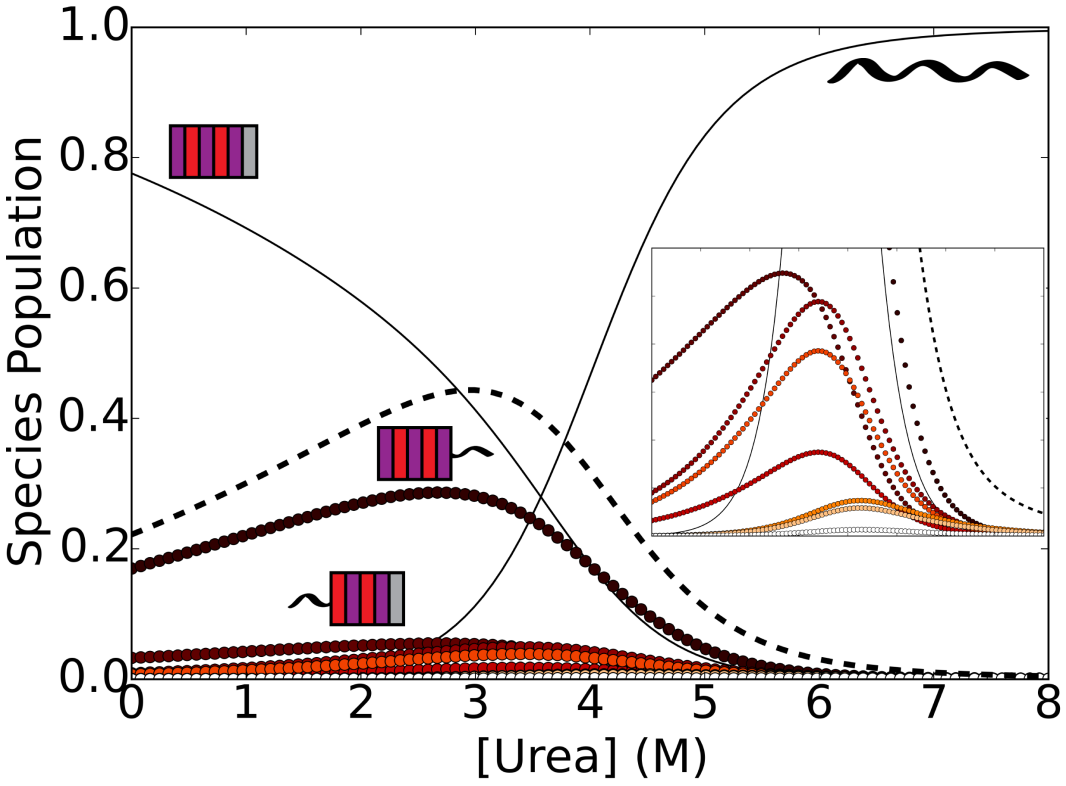


Figure 2.10. Two-dimensional species plot of $B(AB)_2S$. Solid lines represent fully folded and unfolded states, as indicated by their cartoons. Dashed line corresponds to the summation of all microstates with two or more consecutively folded helices, which represent greater than 99% of the total intermediate population. The two most highly populated microstates are plotted, and indicated by their cartoon representations. The inset zooms to show microstate populations with consecutively folded helices.

Relating equilibrium energetics to folding kinetics

Energetics described by an Ising model yield information regarding the equilibrium cooperativity in folding, they say nothing about the rate at which folding occurs. There have been a number of studies, which have found a correlation between stability and folding rate (Aksel and Barrick, 2014; Dao et al., 2015; Torquato et al., 1990; Tripp and Barrick, 2008). Previous kinetic studies of c34PRs have conflicting interpretations. In a 2005 study by Regan and coworkers, increased stability in c34PRs was found to be mainly a result of a slower unfolding rate. However, their data also suggested a folding rate enhancement upon the addition of single repeats (Main et al., 2005). Javadi and Main studied the length dependence of the kinetics of folding and unfolding of a series of c34PRs ranging from 2.5 to 10.5 total repeats (Javadi and Main, 2009). The authors propose two-state like folding for $(AB)_2S$, and multistate folding through stable intermediates for constructs greater than 2.5 repeats.

Although aspects of their data support this conclusion, the data are rather noisy, and their chevron plot folding and unfolding arms are non-linear, making results difficult to interpret. In addition, while the folding arm is non-linear at low concentrations of denaturant, folding rates at higher denaturant suggest a folding rate enhancement with respect to repeat number, though this effect is not accounted for in their model (Javadi and Main, 2009).

Due to the regular architecture of observed in c34PRs, they provide an excellent test of the relationships between stability and folding rate (Aksel and Barrick, 2014). If c34PRs fold via parallel pathways, we would expect a similar rate enhancement upon successive additions of repeats. In addition, our ability to subdivide this system into α-helices provides another variable, namely, to study folding rate as a function of helical additions, as any differential enhancement in folding rate would be due to α-helical stability since the two types of interfaces are of equal energy in c34PRs (Table 2.1). To determine if we could observe linear folding arms, we collected chevron plots of c34PR constructs $(AB)_2S$, $(AB)_3S$, and $(AB)_4S$.

Our initial results on $(AB)_2S$ indicated that better determined chevron plots could be could be generated at lower temperatures (Figure 2.11). We observed a slight roll-over in the unfolding and refolding arms. We collected data for $(AB)_3S$, and $(AB)_4S$ at 10°C which resulted in very clean and reproducible data (Figure 2.12). The refolding arms of these constructs have sharp kinks, consistent with the observations of Javadi and Main. In addition the reduced temperatures make the study of more stable constructs difficult, as solubility of denaturants is decreased, and the stability of these constructs is likely increased. However, we did observe a clear enhancement of the folding rate at denaturant concentrations above this kink, which is consistent with the observation of

parallel or multistate folding with a similar transition state for cANKs (Aksel and Barrick, 2014).

Analysis of fluorescence amplitudes indicates the burst phase intermediate has the same intrinsic fluorescence as the unfolded state (data not shown). While roll-overs are traditionally interpreted as intermediates, there has been evidence they can arise through aggregation (Silow and Oliveberg, 1997; Went et al., 2004). To determine if this occurs with c34PR folding, we collected re-folding traces over a range of concentrations for $(AB)_3S$ (Figure 2.13). The folding rate constants are independent over the protein concentration ranges we examined (1-35 μM) do not show any alteration of the rate constants for folding. This suggests the roll-over is caused by a monomeric intermediate which has an identical fluorescence value to the denatured state. These findings are entirely consistent with those of Javadi and Main.

While I have not taken these kinetic studies further, there have been successful attempts to minimize intermediates through mutation (Wu et al., 2007). If the physical properties of the intermediate involve hydrophobic residues, their role can be measured using 1-anilinonaphthalene-8-sulfonate (ANS), and by substituting them with polar residues (Jones et al., 1994). Javadi and Main were able to detect binding of ANS for the longer (>3.5 repeats) constructs. Therefore, it may prove to

be a successful strategy to substitute hydrophobic residues in c34PRs in an effort to observe linear refolding arms.

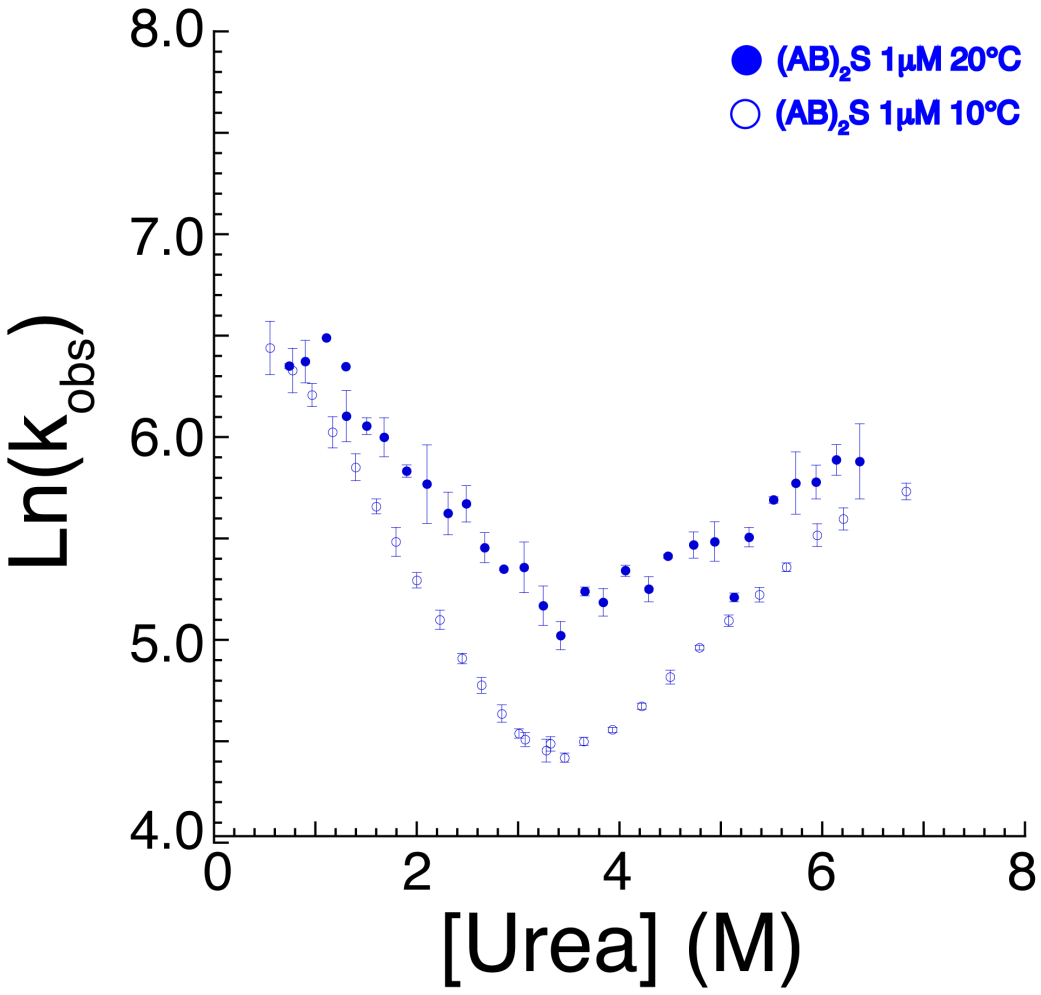


Figure 2.11. Temperature dependence of the folding and unfolding rates of the c34PR $(AB)_2S$. Error bars represent standard errors of fitted rate constants from three or more independent progress curves.

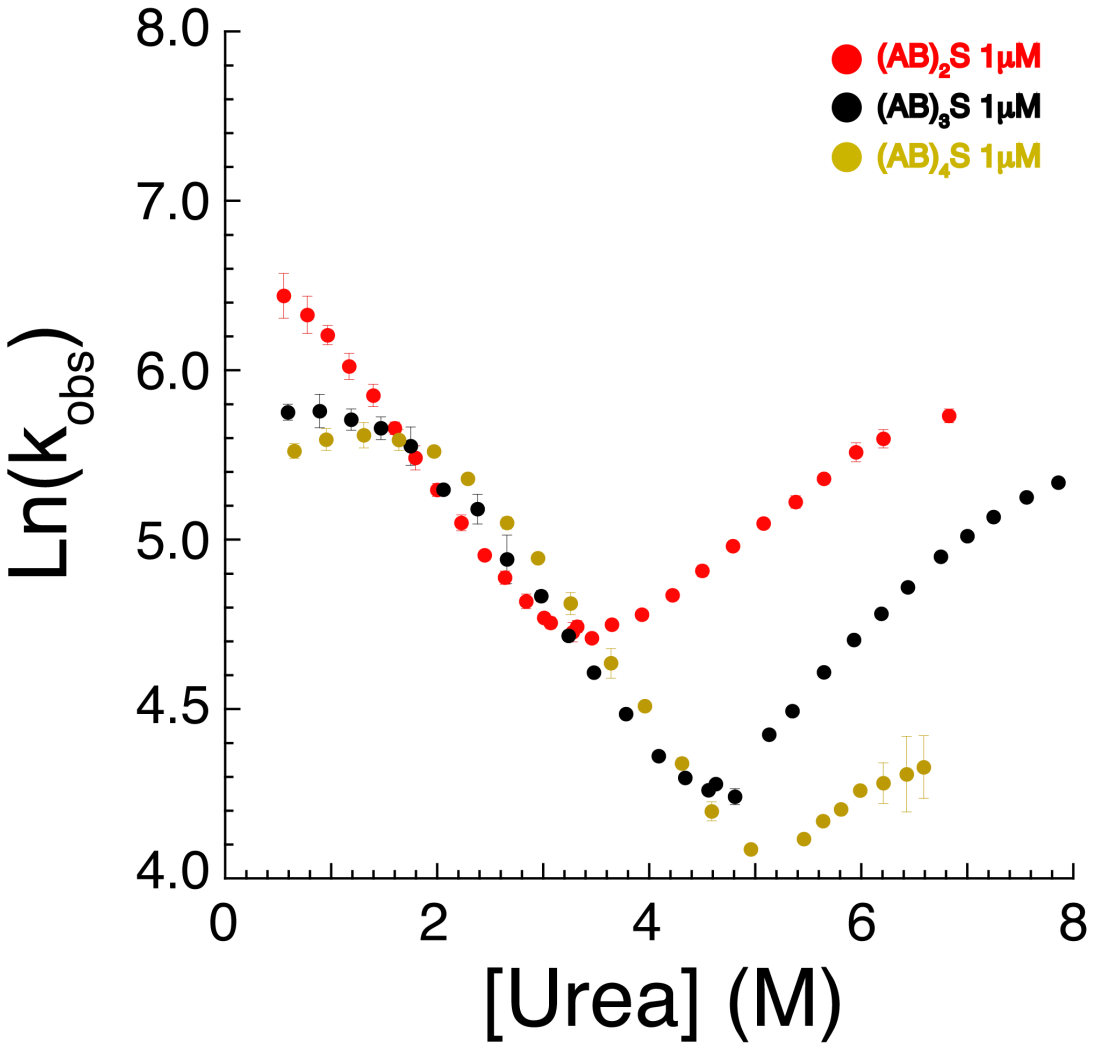


Figure 2.12. Kinetic chevron plots of c34PR constructs $(AB)_2S$, $(AB)_3S$, and $(AB)_4S$. Data were collected at 10°C. Error bars represent standard errors of fitted rate constants from three or more independent shots.

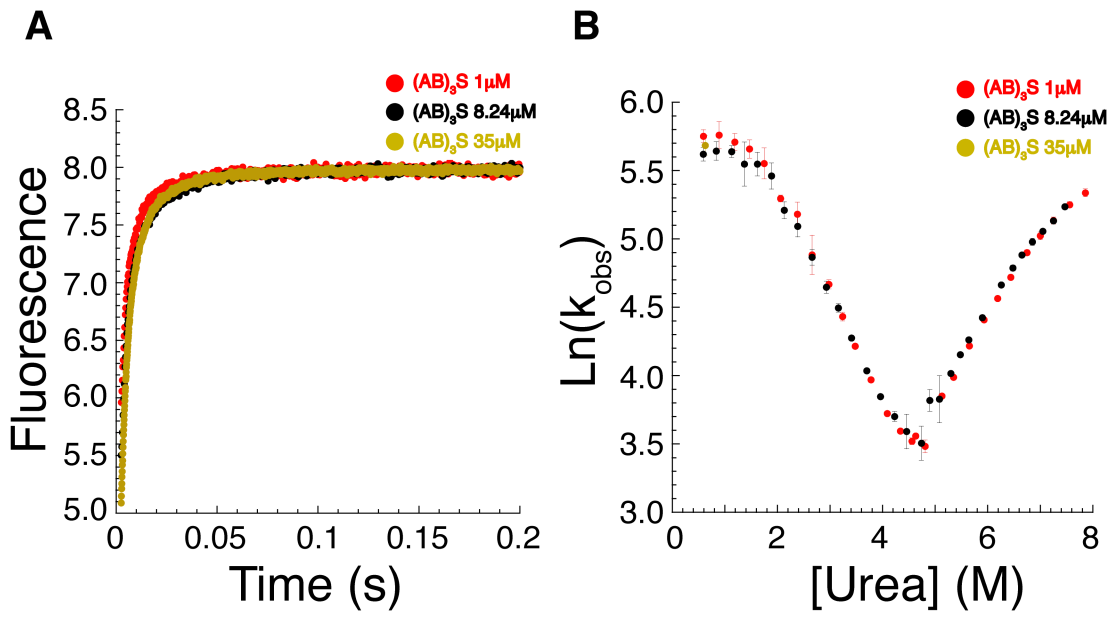


Figure 2.13. Example refolding kinetic traces and chevron plots of c34PR $(AB)_3S$ at multiple concentrations. (A) Refolding kinetic traces at three different concentrations. (B) chevron plots for two concentrations, along with lowest [urea] refolding trace collection at 35 μM. Concentrations ranged from 1-35 μM. Data were collected at 10°C. Error bars represent standard errors of fitted rate constants from three or more independent shots.

2.5 Experimental Procedures

Subcloning, protein expression, and purification

DNA sequences encoding AB, A, B, and S sequences (Figure 1) were cloned by annealing complementary, codon optimized oligonucleotides. Annealed single-repeat/helix cassettes were ligated directly into NdeI and BglII digested pET-15b (Novagen, Madison, WI). BamHI and BglII sites were included in appropriate sequences for repeat assembly as previously described (Aksel et al, 2012).

c34PR constructs were expressed in *Escherichia coli* Rosetta R2* (DE3) cells and purified as described in Chapter 3, in buffer containing 20 mM Na Phosphate, 500 mM NaCl, 25 mM imidazole, pH 7.4, 1mg DNase. Tagged proteins were purified from the supernatant or pellet via Ni-NTA chromatography. Purified proteins were dialyzed extensively into “storage buffer” containing 50 mM Na Phosphate, 150 mM NaCl, pH 6.8, concentrated using an Amicon stirred cell concentrator (EMD Millipore, USA), and flash frozen at -80◦C. Protein concentrations were determined as described by Edelhoch (Edelhoch, 1967).

Circular dichroism spectroscopy

CD measurements were conducted using an Aviv Model 400 CD Spectropolarimeter (Lakewood, NJ) as described in Chapter 3. Far-UV CD

spectra were collected in storage buffer at protein concentrations ranging from 15-35 μM.

Nuclear magnetic resonance spectroscopy

Doubly labeled ^{1}H-^{15}N c34PRs were expressed overnight in M9 minimal medium supplemented with $^{15}NH_4Cl$ (Cambridge Isotope Laboratories, Andover, MA) at 18°C by induction with IPTG and purified as previously described (Marold et al., 2015; see chapter 3). NMR samples contained 50-200μM protein in storage buffer supplemented with 5% D_2O. Heteronuclear HSQC spectra of collected on a Bruker Advance II 600 MHz spectrometer equipped with a cryoprobe, and processed with NMRPipe (Delaglio et al., 1995) at 25°C and displayed using CARA (Keller, 2004).

Analytical Ultracentrifugation

Analytical ultracentrifugation sedimentation velocity (AUC-SV) experiments were performed using a ProteomeLab-equipped Beckman XL-I analytical ultracentrifuge. All AUC experiments were performed on proteins extensively dialyzed into storage buffer. Dilutions were performed using the dialysate. Cell assembly and experimental procedures were conducted as described in Chapter 3.

Urea-induced equilibrium unfolding transitions

Equilibrium unfolding transitions were obtained by monitoring CD at 220 nm in a 1 cm path-length quartz cuvette and conducted as described in Chapter 3. Protein concentrations ranged from 1.0-3.5 μM.

Heteropolymeric nearest-neighbor analysis of c34PR constructs

Nearest-neighbor analysis of c34PR constructs was conducted by generating partition functions (*q*) for each construct (Figures 2.4-2.5). Statistical weights were substituted by exponentiating Boltzmann partial weighted global free energy parameters (ΔG_A, ΔG_B, ΔG_S, $\Delta G_{Ai:Bi}$, $\Delta G_{Bi:Ai+1}$). By differentiation of *q* with respect to $\Delta G_{A,B,S}$, the fraction of folded helices (Φ) is generated:

$$\Phi_p = \frac{1}{n*\mathrm{q}_p} \times \sum_{i\,=\,helix\ types} e^{(\Delta G_i - (m_i*[x])/RT)} * \frac{\partial q_p}{\partial K_i}$$

(2.14)

where p designates the individual protein, n represents the number of helices, m_i represents the denaturant dependence on the intrinsic free energy, and x is the concentration of denaturant. This term is then used in a simple function to connect to an experimental observable:

$$Y_{obs} = \Big(\big((a * [x]) + b\big) * \Phi\Big) + \Big(\big((c * [x]) + d\big) * (1 - \Phi)\Big)$$

(2.15)

where a , b and c , d are individual native and denatured baseline parameters, respectively. Parameter optimization was obtained through least-squares minimization routine using a suite of Python programs I developed. To enhance clarity and interpretation, the data and fitted curves were baseline adjusted using the best-fit baseline parameters using the equation:

$$Y_{adjusted} = \frac{\Big(Yobs - \big((a*x)+b\big)\Big)}{\Big(\big((c*x)+d\big) - \big((a*x)+b\big)\Big)}$$

(2.16)

and are shown in Figure 2.3. Uncorrected data and fits are shown in S2.3.

Stopped-flow fluorescence kinetics

Fluorescence detected stopped-flow kinetic measurements were performed on an Applied Photophysics (Leatherhead, England) SX.18MV-R stopped-flow fluorometer. Prior to measurements, all samples and buffer were de-gassed for two or more hours. Final protein concentrations

ranged from 1 to 35 μM. Reactions were monitored by excitation at 280 nm, recording emission using a 320 nm cutoff filter. All experiments were performed in storage buffer at 10°C.

Analysis of kinetic traces was performed by fitting the following equation using non-linear least squares methods:

$$Y_{obs} = Y_{\infty} + \sum_i \Delta Y_i * e^{-k_i t} \quad (2.17)$$

here, Y_{∞} is the fluorescence signal at equilibrium, ΔY_i represents the fluorescence change for the i^{th} kinetic phase, and k_i is the rate constant. To determine the correct number of phases to each unfolding trace, I developed a Python program, KinetiChevron, to rapidly fit and display multiple kinetic models side-by-side. An example of the output for model interpretation is shown in S2.7.

2.6 Supplemental information

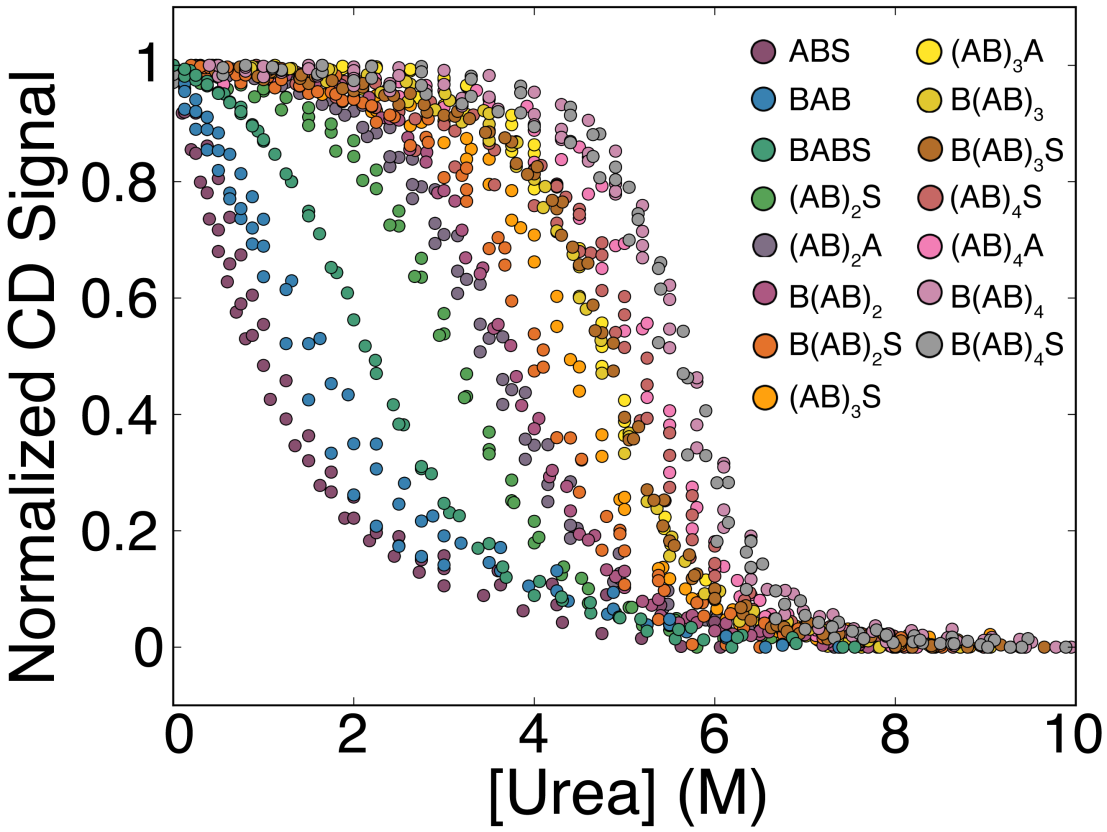
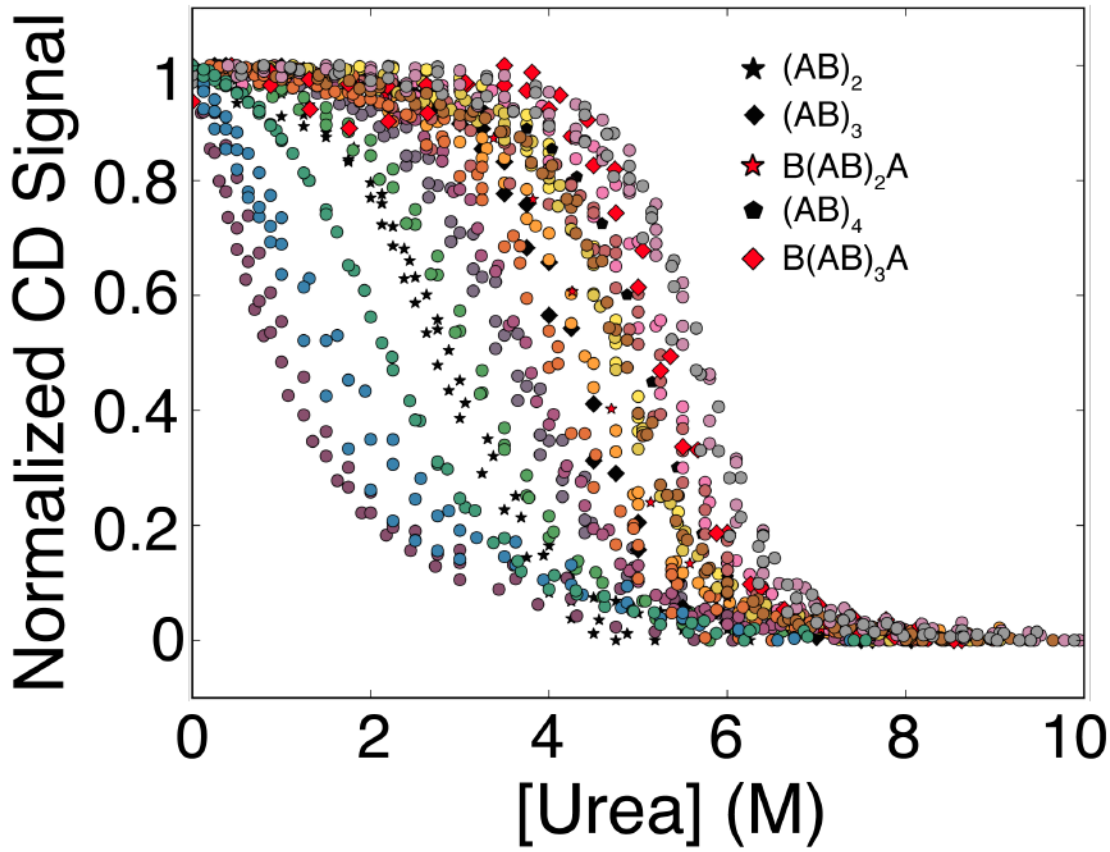
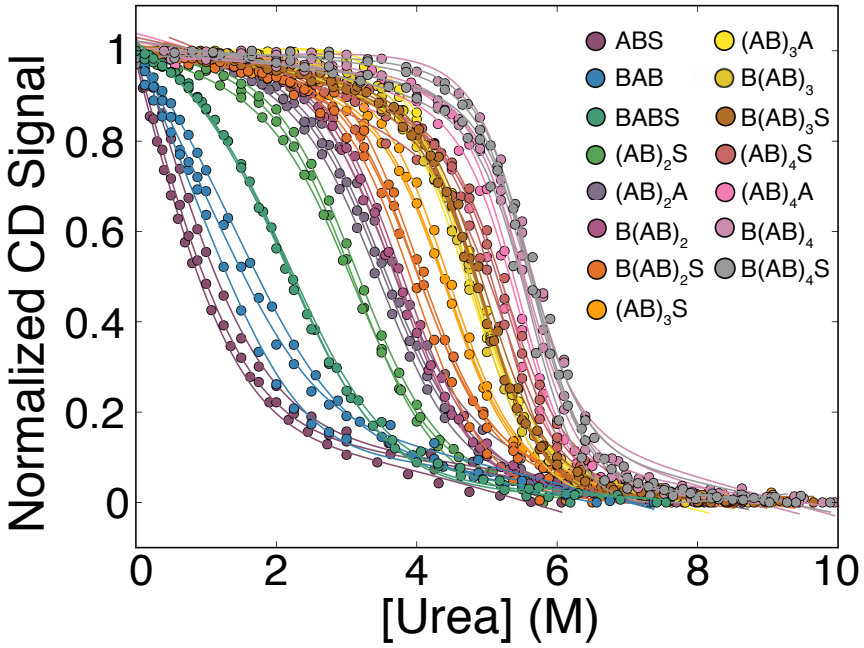


Figure S2.1. Normalized CD signal of c34PR constructs studied. Construct data are colored the same as in Figure 2.2. These data represent the raw normalized CD signal and are not baseline adjusted or fit to yield fraction folded values.



S2.2. Normalized CD signal of c34PR constructs, including those which result in self-association. These data represent the raw normalized CD signal and are not baseline adjusted or fit to yield fraction folded values. Colors for constructs which do not self-associate are colored the same as in Figure 2.2. Colors and symbols for constructs which do self-associate are displayed in the legend. There are two experiments displayed for $(AB)_2$ (the construct with the lowest stability of those in the legend).



S2.3. Model M1 heteropolymeric Ising model fit. Colors are the same as in Figure 2.2 These data are not baseline adjusted to yield fraction folded values. The Ising free energy values and errors in Table 2.1 are determined from this fit.

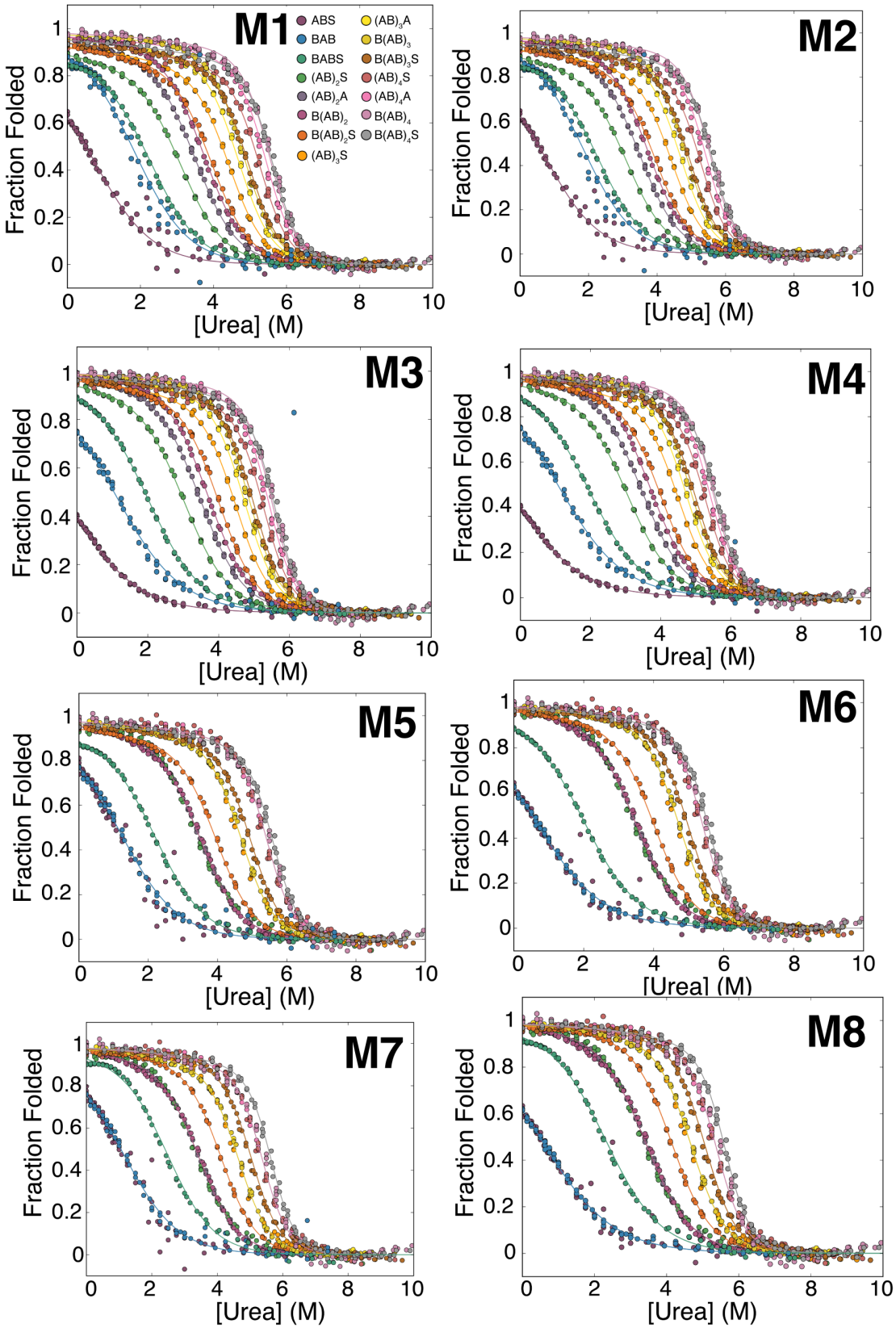

M1
ABS
BAB
BABS
(AB)2S
(AB)2A
B(AB)2
B(AB)2S
(AB)3S
(AB)3A
B(AB)3
B(AB)3S
(AB)4S
(AB)4A
B(AB)4
B(AB)4S
M2
M3
M4
M5
M6
M7
M8
Fraction Folded
[Urea] (M)

Figure S2.4. Ising fits of models M1-M8. Colors are the same as in Figure 2.2. The model indicators M1-M8 correspond to the ones listed in Tables 2.1, 2.2, and 2.3.

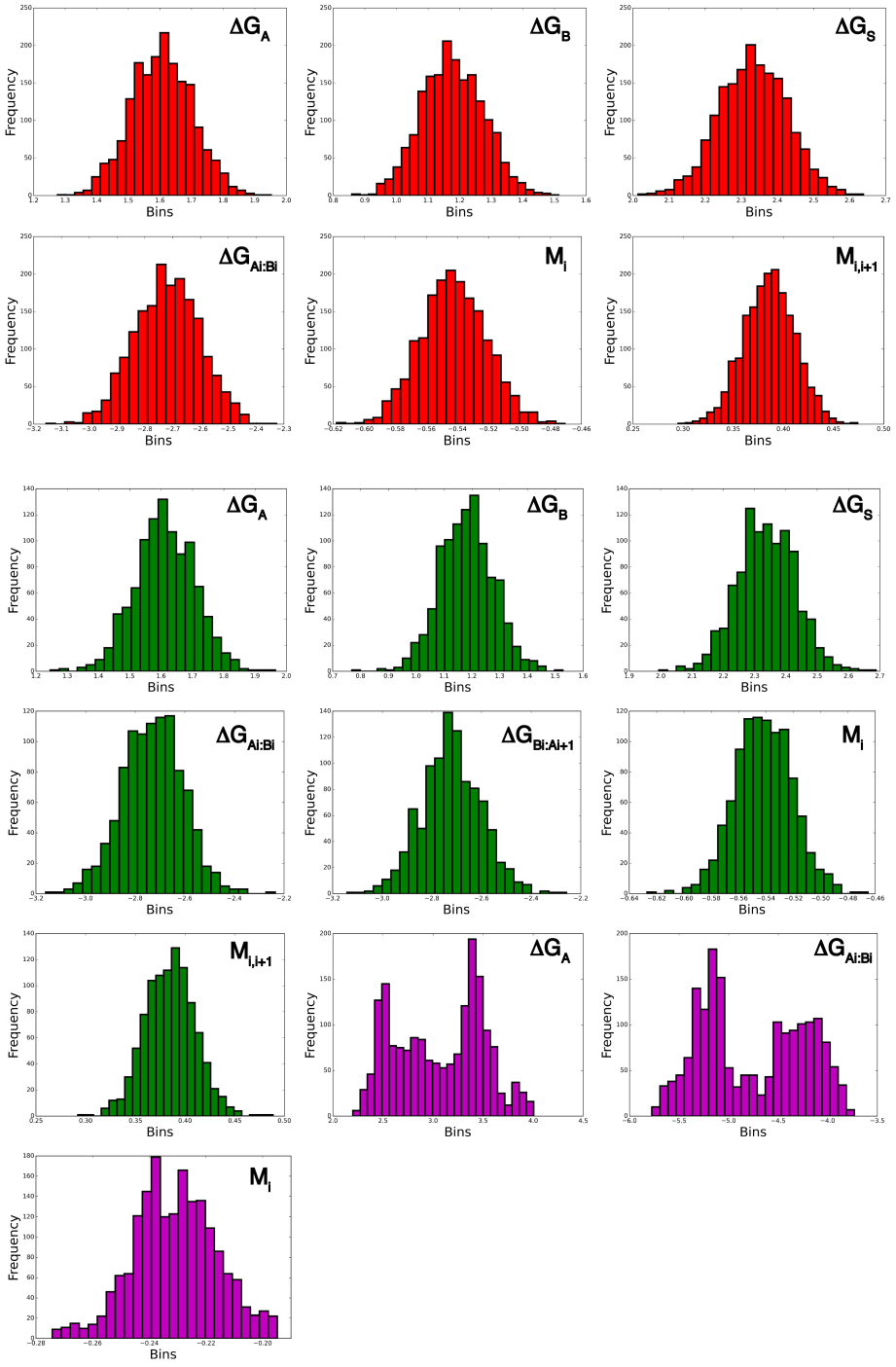


Figure S2.5. c34PR bootstrapped error distributions. Distributions are shown for 1800 bootstrap iterations for models M1 (red), M2 (green), and M8 (magenta).

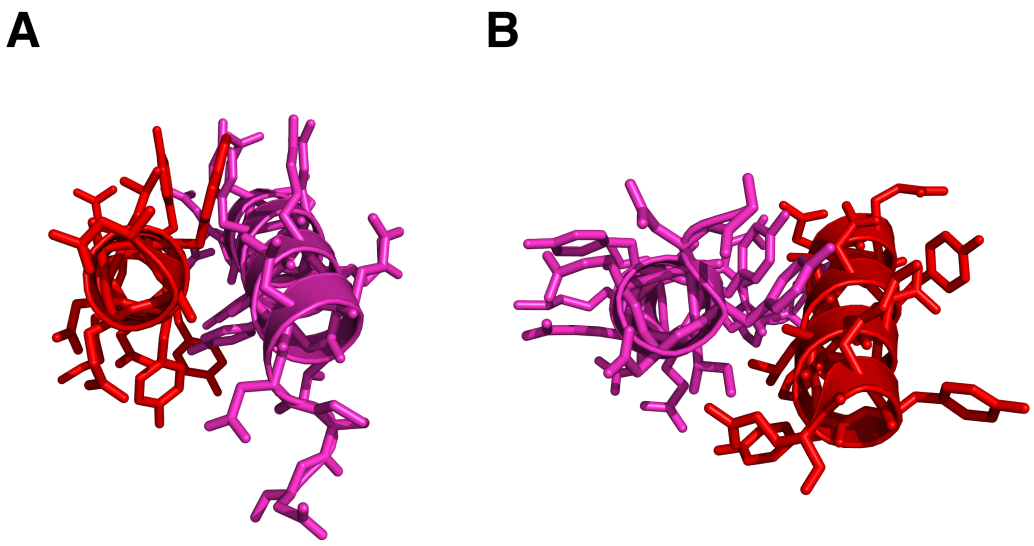


Figure S2.6. Helical packing interfaces in c34PRs. (A) Ai:Bi helical interface. (B) B_i:$A_{i,i+1}$ helical interface. Coordinates were obtained from the structure 1NA0.

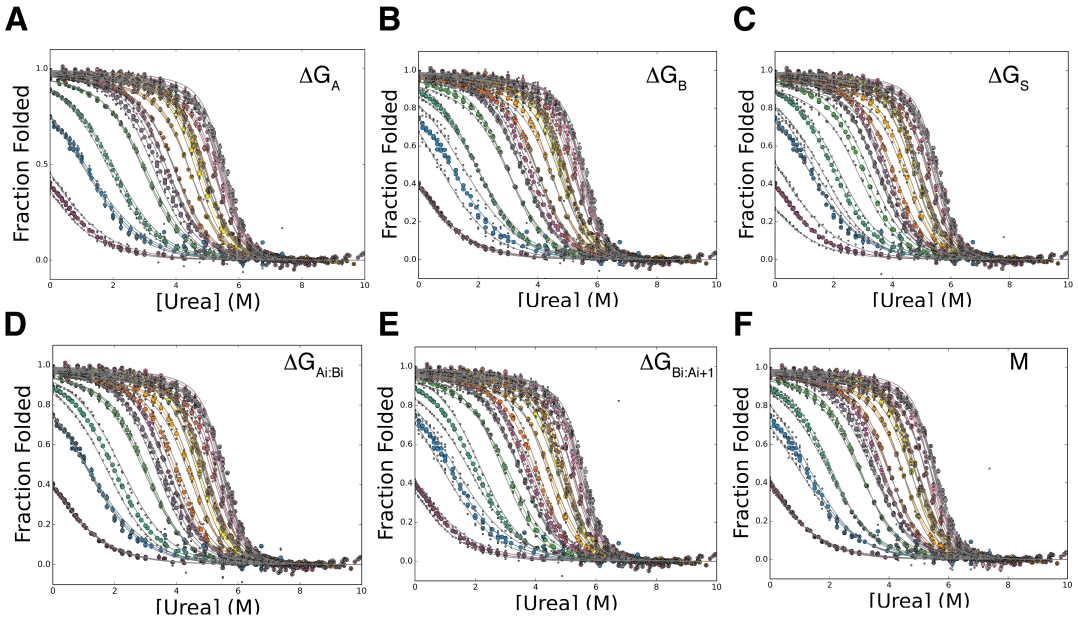


Figure S2.7. F-statistic-derived 95% confidence interval fits of c34PRs using model M3. Experimental data for constructs are displayed as solid circles, and colored the same as in Figure 2.2. The 95% confidence limits for each global parameter are shown as smaller grey dots (transformed experimental data to fraction folded) and lines (fit). (A-F) 95% confidence limit fits for ΔGA, ΔGB, ΔGS, $\Delta GA_{i:Bi}$, $\Delta GB_{i:Ai,i+1}$, and m_i, respectively.

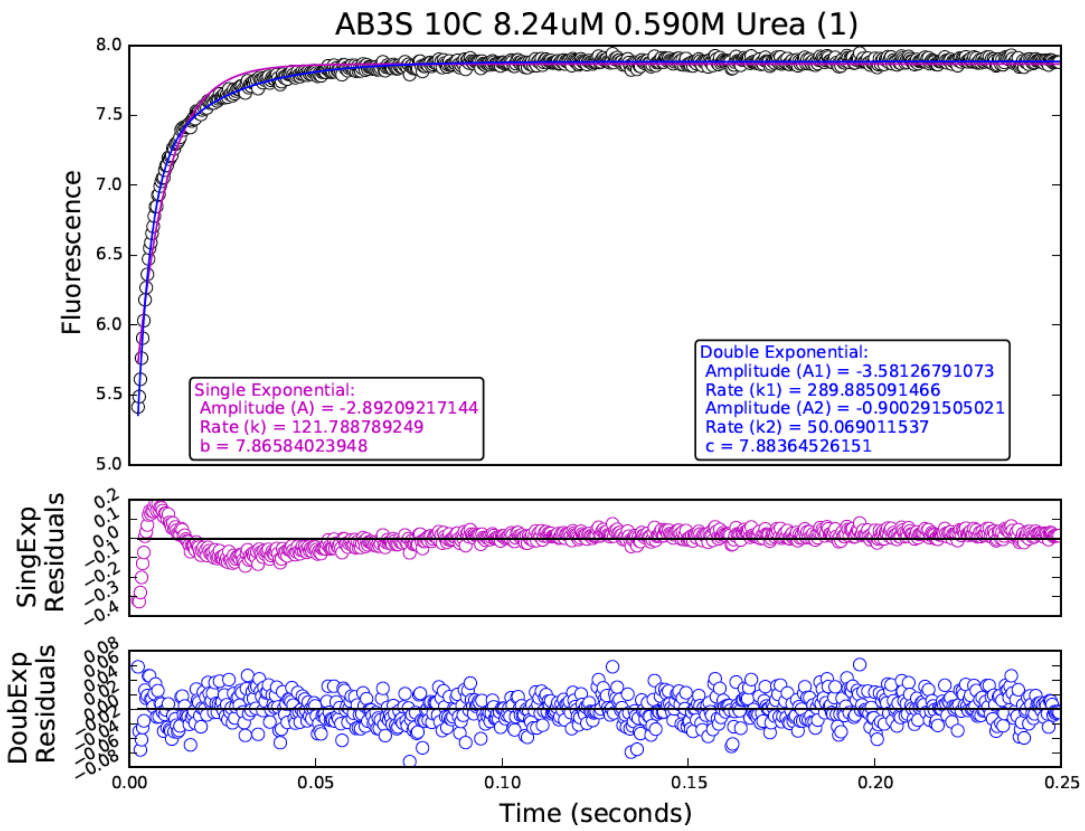


Figure 2.8. Example output of the program KinetiChevron. Kinetic traces are displayed in the upper panel fitted to single (magenta) and double (blue) exponential models. Residuals for each are shown below the trace, on the same x-axis.

2.7 References

Aksel, T., and Barrick, D. (2009). Chapter 4 Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models. In Methods in Enzymology, (Elsevier), pp. 95–125.

Aksel, T., and Barrick, D. (2014). Direct Observation of Parallel Folding Pathways Revealed Using a Symmetric Repeat Protein System. Biophysical Journal *107*, 220–232.

Aksel, T., Majumdar, A., and Barrick, D. (2011). The Contribution of Entropy, Enthalpy, and Hydrophobic Desolvation to Cooperativity in Repeat-Protein Folding. Structure *19*, 349–360.

Baldwin, R.L. (2007). Energetics of Protein Folding. Journal of Molecular Biology *371*, 283–301.

Ballerini, M., Cabibbo, N., Candelier, R., Cavagna, A., Cisbani, E., Giardina, I., Lecomte, V., Orlandi, A., Parisi, G., Procaccini, A., et al. (2008). Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. Proceedings of the National Academy of Sciences *105*, 1232–1237.

Carrion-Vazquez, M., Oberhauser, A. F., Fowler, S. B., Marszalek, P. E., Broedel, S. E., Clarke, J. and Fernandez, J. M. (1999). Mechanical and chemical unfolding of a single protein: a comparison. Proc. Natl Acad. Sci. USA 96, 3694 – 3699.

Chatelier, R. C. (1987). Indefinite isoenthalpic self-association of solute molecules. *Biophysical chemistry*, *28*(2), 121-128.

Cortajarena, A.L., and Regan, L. (2011). Calorimetric study of a series of designed repeat proteins: Modular structure and modular folding. Protein Science *20*, 336–340.

Courtemanche, N., and Barrick, D. (2008). Folding thermodynamics and kinetics of the leucine-rich repeat domain of the virulence factor Internalin B. Protein Science *17*, 43–53.

D'Andrea, L. (2003). TPR proteins: the versatile helix. Trends in Biochemical Sciences *28*, 655–662.

Dao, T.P., Majumdar, A., and Barrick, D. (2014). Capping motifs stabilize the leucine-rich repeat protein PP32 and rigidify adjacent repeats:

Roles of Caps in the Folding of the LRR Protein PP32. Protein Science *23*, 801–811.

Dao, T.P., Majumdar, A., and Barrick, D. (2015). Highly polarized C-terminal transition state of the leucine-rich repeat domain of PP32 is governed by local stability. Proceedings of the National Academy of Sciences *112*, E2298–E2306.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A.D. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. Journal of Biomolecular NMR *6*, 277–293.

Doig, A.J., Andrew, C.D., Cochran, D.A., Hughes, E., Penel, S., Sun, J.K., Stapley, B.J., Clarke, D.T., and Jones, G.R. (2001). Structure, stability and folding of the-helix. In Biochem. Soc. Symp, pp. 95–110.

Doyle, M.L., Speros, P.C., LiCata, V.J., Gingrich, D., Hoffman, B.M., and Ackers, G.K. (1991). Linkage between cooperative oxygenation and subunit assembly of cobaltous human hemoglobin. Biochemistry *30*, 7263–7271.

Efron, B., and Tibshirani, R. (1993). An introduction to the bootstrap (New York: Chapman & Hall).

Englander, S.W., and Mayne, L. (2014). The nature of protein folding pathways. Proceedings of the National Academy of Sciences *111*, 15873–15880.

Englander, S.W., Mayne, L., and Krishna, M.M.G. (2007). Protein folding and misfolding: mechanism and principles. Quarterly Reviews of Biophysics *40*.

Ferreiro, D.U., and Wolynes, P.G. (2008). The capillarity picture and the kinetics of one-dimensional protein folding. Proceedings of the National Academy of Sciences *105*, 9853–9854.

Guerois, R., and López de la Paz, M. (2006). Protein design: methods and applications (Totowa, N.J.: Humana Press).

Hildenbrandt, H., Carere, C., and Hemelrijk, C.K. (2010). Self-organized aerial displays of thousands of starlings: a model. Behavioral Ecology *21*, 1349–1359.

Hongbin, L. and Fernandez, J.M. (2003). Mechanical design of the first proximal Ig domain of human cardiac titin revealed by single molecule force spectroscopy. Journal of Molecular Biology 334, 75-86.

Hu, W., Walters, B.T., Kan, Z.-Y., Mayne, L., Rosen, L.E., Marqusee, S., and Englander, S.W. (2013). Stepwise protein folding at near amino acid resolution by hydrogen exchange and mass spectrometry. Proceedings of the National Academy of Sciences *110*, 7684–7689.

Ising, E. (1925). Title Unavailable. Z. Physik 31,253.

Johnson, M.L. (2008). Nonlinear Least-Squares Fitting Methods. In Methods in Cell Biology, (Elsevier), pp. 781–805.

Kajander, T., Cortajarena, A.L., Main, E.R.G., Mochrie, S.G.J., and Regan, L. (2005). A New Folding Paradigm for Repeat Proteins. Journal of the American Chemical Society *127*, 10188–10190.

Kajava, A.V. (2001). Review: Proteins with Repeated Sequence—Structural Prediction and Modeling. Journal of Structural Biology *134*, 132–144.

Karpenahalli, M.R., Lupas, A.N., and Söding, J. (2007). TPRpred: a tool for prediction of TPR-, PPR-and SEL1-like repeats from protein sequences. BMC Bioinformatics *8*, 2.

Kiefhaber, T., and Baldwin, R.L. (1995). Intrinsic Stability of Individual α Helices Modulates Structure and Stability of the Apomyoglobulin Molten Globule Form. Journal of Molecular Biology *252*, 122–132.

Kloss, E., Courtemanche, N., and Barrick, D. (2008). Repeat-protein folding: New insights into origins of cooperativity, stability, and topology. Archives of Biochemistry and Biophysics *469*, 83–99.

Main, E., Lowe, A., Mochrie, S., Jackson, S., and Regan, L. (2005a). A recurring theme in protein engineering: the design, stability and folding of repeat proteins. Current Opinion in Structural Biology *15*, 464–471.

Main, E.R., Stott, K., Jackson, S.E., and Regan, L. (2005b). Local and long-range stability in tandemly arrayed tetratricopeptide repeats. Proceedings of the National Academy of Sciences of the United States of America *102*, 5721–5726.

Main, E.R.G., Xiong, Y., Cocco, M.J., D'Andrea, L., and Regan, L. (2003). Design of Stable α-Helical Arrays from an Idealized TPR Motif. Structure *11*, 497–508.

Mello, C.C., and Barrick, D. (2004). An experimentally determined protein folding energy landscape. Proceedings of the National Academy of Sciences of the United States of America *101*, 14102–14107.

Mor, A., Haran, G., and Levy, Y. (2008). Characterization of the unfolded state of repeat proteins. HFSP Journal *2*, 405–415.

Mosavi, L.K., Minor, D.L., and Peng, Z. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. Proceedings of the National Academy of Sciences *99*, 16029–16034.

Motlagh, H.N., and Hilser, V.J. (2012). Agonism/antagonism switching in allosteric ensembles. Proceedings of the National Academy of Sciences *109*, 4134–4139.

Naganathan, A.N., Perez-Jimenez, R., Muñoz, V., and Sanchez-Ruiz, J.M. (2011). Estimation of protein folding free energy barriers from calorimetric data by multi-model Bayesian analysis. Physical Chemistry Chemical Physics *13*, 17064.

Sikorski, R.S., Michaud, W.A., Wootton, J.C., Boguski, M.S., Connelly, C., and Hieter, P. (1991). TPR Proteins as Essential Components of the Yeast Cell Cycle. Cold Spring Harbor Symposia on Quantitative Biology *56*, 663–673.

Silow, M., and Oliveberg, M. (1997). Transient aggregates in protein folding are easily mistaken for folding intermediates. Proceedings of the National Academy of Sciences *94*, 6084–6086.

Sontag CA, Stafford WF, Correia JJ. A comparison of weight average and direct boundary fitting of sedimentation velocity data for indefinite polymerizing systems. Biophys Chem. 2004;108:215–30.

Sosnick, T.R., and Barrick, D. (2011). The folding of single domain proteins—have we reached a consensus? Current Opinion in Structural Biology *21*, 12–24.

Stanley, H.E. Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, July, 1987. ISBN-10: 0195053168, ISBN-13: 9780195053166.

Torquato, S., Lu, B., and Rubinstein, J. (1990). Nearest-neighbor distribution functions in many-body systems. Physical Review A *41*, 2059.

Tripp, K.W., and Barrick, D. (2007). Enhancing the Stability and Folding Rate of a Repeat Protein through the Addition of Consensus Repeats. Journal of Molecular Biology *365*, 1187–1200.

Tripp, K.W., and Barrick, D. (2008). Rerouting the Folding Pathway of the Notch Ankyrin Domain by Reshaping the Energy Landscape. Journal of the American Chemical Society *130*, 5681–5688.

Viscido, S.V., Parrish, J.K., and Grünbaum, D. (2005). The effect of population size and number of influential neighbors on the emergent properties of fish schools. Ecological Modelling *183*, 347–363.

Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K., and Plückthun, A. (2008). Folding and Unfolding Mechanism of Highly Stable Full-Consensus Ankyrin Repeat Proteins. Journal of Molecular Biology *376*, 241–257.

CHAPTER 3

A naturally occurring repeat protein with high internal sequence identity defines a new class of TPR-like proteins

3.1 Abstract

Linear repeat proteins often have high structural similarity and low (~25%) pairwise sequence identities (PSI) among modules. We identified a unique *Podospora anserina* (*Pa*) sequence with tetratricopeptide repeat (TPR) homology, which contains longer (42 residue) repeats (42PRs) with an average PSI >91%. We determined the crystal structure of five tandem *Pa* 42PRs to 1.6 Å, and examined the stability and solution properties of constructs containing three to six *Pa* 42PRs. Compared to 34-residue TPRs (34PRs), *Pa* 42PRs have a one-turn extension of each helix, and bury more surface area between helices. Equilibrium unfolding transitions become more stable and sharper as *Pa* 42PRs are added, suggesting a higher level of cooperativity in folding compared to consensus 34PRs (c34PRs). These results demonstrate the versatility of the TPR motif to length variation, and provide a basis to understand the effects of helix

length on intrinsic/interfacial stability using nearest-neighbor (Ising) statistical thermodynamic models.

3.2 Introduction

Linear repeat proteins consist of arrays of a common structural motif, typically 20-40 residues in length. Adjacent motifs pack together to create elongated structures with superhelical properties defined by geometric relationships between units (Kloss et al., 2008; Main et al., 2005; Kajava, 2002; Kobe and Kajava, 2000). One such motif is the tetratricopeptide repeat (TPR), a 34-residue sequence motif found in a wide range of proteins from all three kingdoms of life.

TPR domains mediate protein-protein interactions. Although TPR sequences and functions vary widely, repeats have nearly identical geometries (D'Andrea and Regan, 2003). The structure of TPR motifs consists of two anti-parallel α-helices, termed "A" and "B", which stack at an angle of ~160° (Blatch and Lässle, 1999). The structure and folding of tandem 34 residue TPR domains has been studied extensively using a series of consensus repeats, in which each repeat has the same sequence, based on multiple sequence alignments (Main et al., 2003; Kajander et al., 2005; Cortajarena and Regan, 2011). The application of consensus design methods to repeat proteins (Binz et al., 2003; Main et

al., 2003; Mosavi et al., 2002; Parmeggiani et al., 2008; Urvoas et al., 2010), which is usually based on hidden Markov models (HMMs), highlights conserved residues of each motif. However, HMM programs are infrequently used to generate new motifs (Frith et al., 2008), and length variations in aligned sequences are therefore reduced to insertion and deletion probabilities within HMMs. This has the potential to mask significant length differences among distinct motif subfamilies.

One particularly interesting aspect of the TPR motif, compared to other linear repeat motifs, is the diversity of its repeat sequence. The Pfam 27.0 (Finn et al., 2014) TPR superfamily contains over 100 family members. Of these, 21 family members are classified TPR_1 through TPR_21. Although some family members have very similar HMM logos and consensus sequences (e.g. TPR_1 and TPR_2), other families differ in length and composition (length range: 26-280 residues). Though some of the longer families result from a classification of tandem repeats as a single motif (presumably due to high similarity between nonadjacent repeats), there is considerable length variation among families representing single repeats. This differs from other linear helical repeats such as ankyrin repeats, where sequence lengths are more tightly distributed (~33 residues/repeat).

A striking example of repeat length variation in TPRs can be found in sequences classified as TPR_10. The sequences in this family are 42

residues in length, as opposed to the founding 34 residue motif (Sikorski et al., 1990). Owing to the length variation observed in TPR sequences, we adopt a nomenclature that better reflects motif length: nPRs (the name TPR derives from the tetratrico prefix, meaning thirty-four; the "T" (for "tetra", four) cannot capture variation in the tens digit). Here, n corresponds to the number of residues in a single repeat. For example, we refer to 42 residue nPR motifs as 42PRs, and 34 residue motifs as 34PRs.

There is little high-resolution structural information for 42PRs. The closest structural homologs are the TPR domains of human kinesin light chain (hKLC) isoforms 1 and 2, which were solved to 2.8 and 2.75 angstroms, respectively (Zhu et al., 2012). Some of the TPRs in hKLC1 and hKLC2 belong to TPR_10, although there are no consecutive repeats classified as TPR_10, perhaps due to the low identity between repeats. This limits an understanding of the structural features defining repeats belonging to 42PRs.

To explore the structural and thermodynamic implications of this new class of extended repeat sequences, we identified and characterized a rather unusual 42PR array from the *Podospora anserina* (*Pa*) genome sequence (Espagne et al., 2008), containing 15 tandem 42PRs of nearly complete identity. We used the repeats in this sequence to design central (AB) and capping (N_AB and AC_B) 42PRs to create a series of constructs of the type $N_AB(AB)_xAC_B$ (*Pa* 42PRs). Here, the subscript x signifies the

number of central AB units, ranging from one to four. We find $N_AB(AB)_xAC_B$ constructs to be soluble, stable, and predominantly (>95%) monomeric below 10μM. We determined the X-ray structure of $N_AB(AB)_3AC_B$ to 1.6Å. The structure reveals a five-repeat *Pa* 42PR right-handed superhelix, with longer A and B helices compared to canonical 34PRs. We find *Pa* 42PRs to be significantly less stable, yet more cooperative, than consensus 34PRs (c34PRs) of equivalent repeat number.

3.3 Results

Identification of a new class of 42PRs

Comparison of the lengths of TPR families 1-21 in Pfam 27.0 revealed two well-represented repeat sequence lengths: 34 and 42 residues. We found many of the 42PR containing sequences to have a rather high average pairwise identity among repeats (internal sequence identity). This differs from the internal sequence identity in 34PRs, and the vast majority of other repeat protein motifs, which is typically only ~25%.

To better define the shared and unique sequence characteristics between 42PRs and 34PRs, we generated HMM sequence logos using seed sequences from Pfam 27.0. Seed sequences TPR_1 and TPR_2 were combined (924 total) to create a 34PR HMM (Figure 3.1A), and 291

TPR_10 seed sequences were used to create a 42PR HMM (Figure 3.1B). Alignment of the 42PR and 34PR HMM logos shows conserved sequence features along the entire 34PR AB unit (Figure 3.1). Residues 2-31 in 34PRs and 6-35 in 42PRs show high sequence identity and have identical spacing between conserved positions. In this alignment, the 42PR sequence logo contains four-residue extensions on the N- and C-termini (residues 1-4 and 34-38 in the 42PR HMM, Figure 3.1B), which could conceivably extend the A and B helices of the 34PR motif.

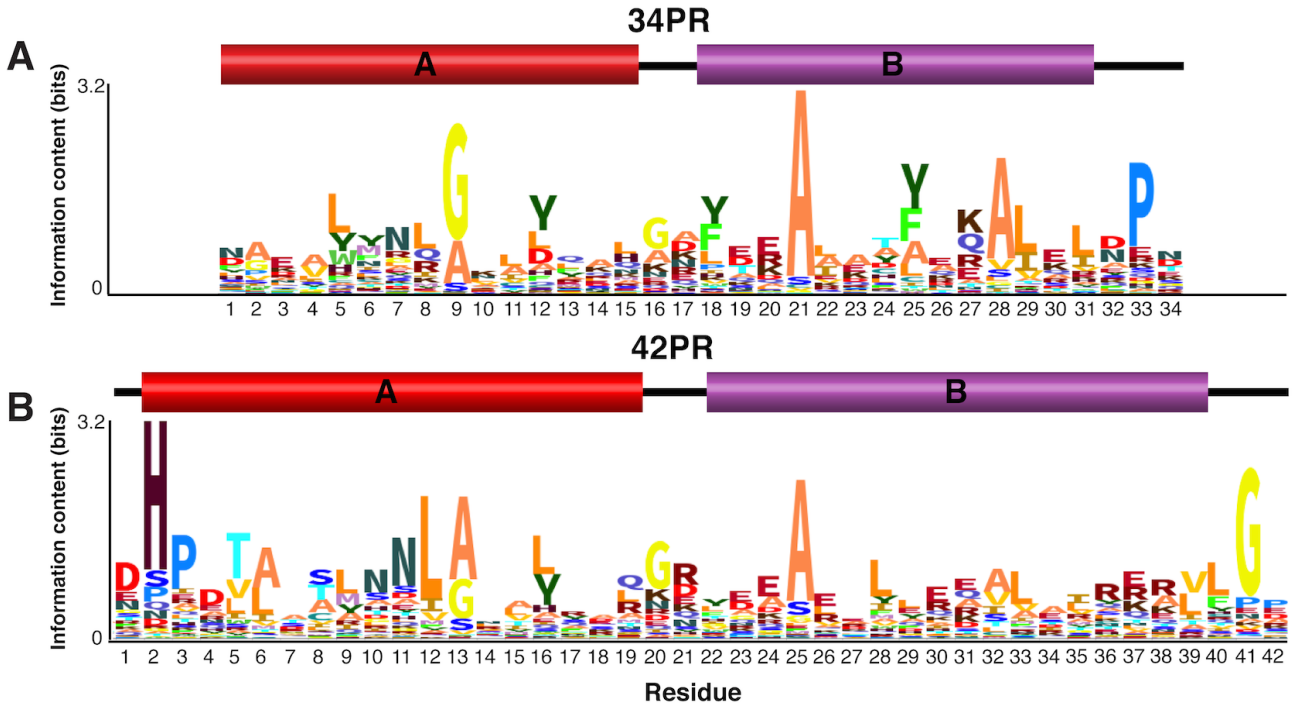


Figure 3.1. HMM logos and helix definitions of 34PR and 42PR sequences. (A) 34PR HMM sequence logo. (B) 42PR HMM sequence logo. A and B-helix boundaries are from the structures 1NA0 (c34PR) and 4Y6W (42PR). Logos were generated using Skylign (Wheeler et al., 2014) and aligned manually in the central region (residues 5-29 in the 34PR logo), which contains the greatest similarity between the two motifs.

Design of *Podospora anserina* repeats

We used the 42PR HMM described above as a guide to search for sequence members of this family. We identified a 42PR-containing ORF, *Pa_6_8860*, present in the genome of the fungus *Podospora anserina*. The predicted domain architecture of *Pa_6_8860* consists of an N-terminal partial NB-ARC domain (van Ooijen et al., 2008), followed by a region containing modest similarity to the heptad repeats in kinesin light chains (Cyr et al., 1991). The remainder of *Pa_6_8860* encodes 15 putative 42PRs (Figure 3.2). The 42PRs of *Pa_6_8860* are an extreme example of high internal sequence identity, with >91% average pairwise identity.

To characterize the structure and stability of this unique *Pa* 42PR, and to compare it with the more common 34PR family, we designed constructs to express arrays of a core AB repeat (*Pa* AB; Figure 3.2B and 3.2C). Although proteins constructed solely from *Pa* AB units expressed well, these polypeptides were insoluble and could not be characterized. Thus, polar substitutions to putative solvent exposed hydrophobic residues were made to the N-terminal A and C-terminal B helices to create capping repeats N_AB and AC_B (Figure 2B). This approach has been successful in promoting solubility in other repeat protein studies (Main et al. 2003; Wetzel et al. 2008; Aksel et al., 2011). We limited substitutions to four surface-exposed sites to minimize structural and energetic perturbations.

We used these capping repeats, along with unmodified AB units, to create constructs of the type $N_AB(AB)_xAC_B$, where the subscript x represents the number of AB core repeats, ranging from one to four (Figure 3.2C). We were able to express and purify these capped constructs with reasonable yields for biophysical characterization.

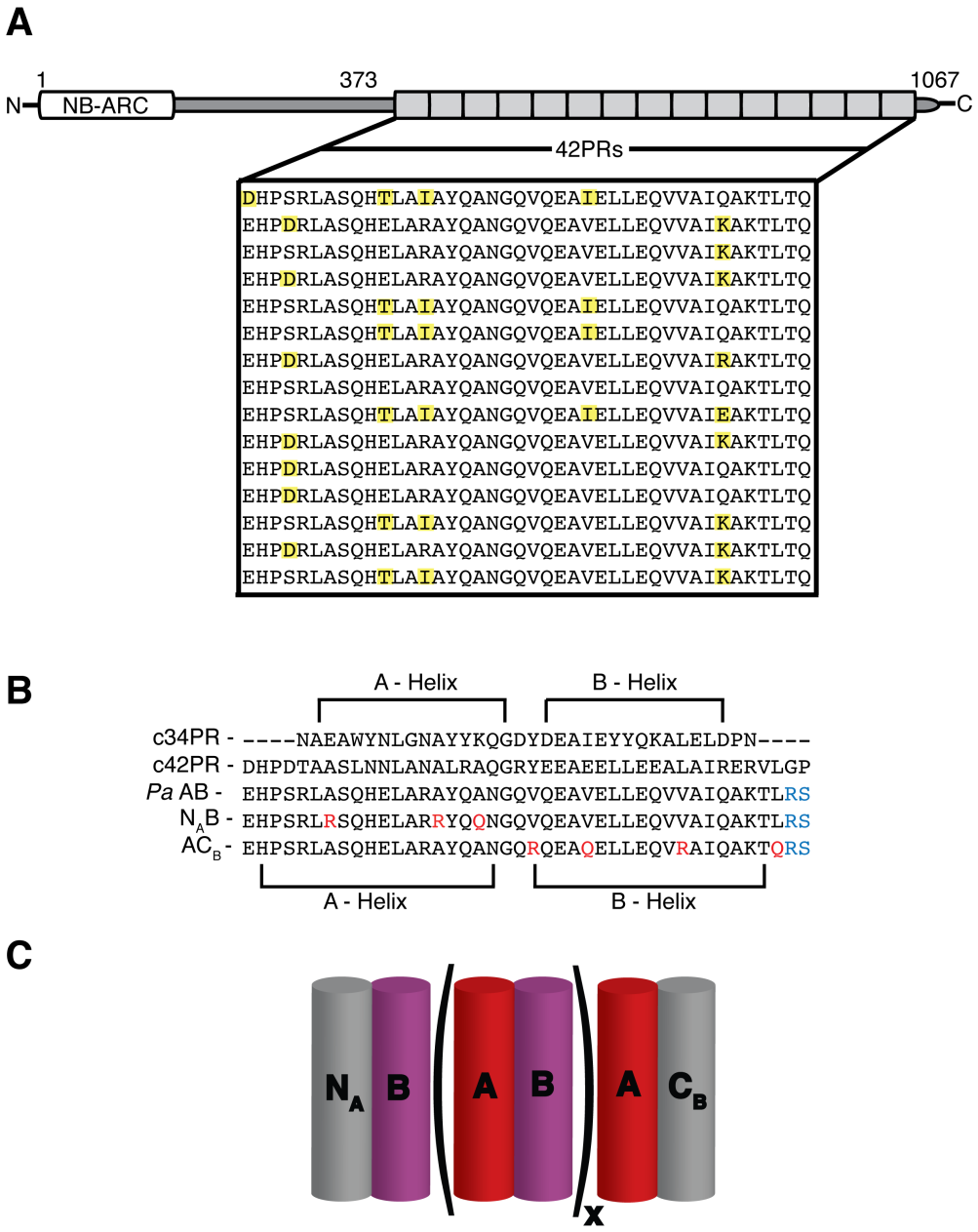


Figure 3.2. Sequence features of *Podospora anserina* Pa_6_8860 ORF and *Pa* 42PR repeat design. (A) Predicted domain organization of *Pa_6_8860*. The predicted 42PR sequence motifs are labeled as grey boxes. Sequence positions in the 15 central 42PRs that differ from the

derived consensus after alignment are highlighted in yellow. (B) Designed *Podospora anserina* 42PR sequence (*Pa* AB) from the alignment in (A). Capping sequences N_AB and AC_B were created by substitution of non-polar residues for polar residues (red) to promote solubility. RS substitutions (blue) resulted from cloning of repeat arrays. The consensus 34PR sequence (Main et al., 2003), c34PR, is aligned as in Figure 3.1. A and B-helix boundaries for the 34PR sequence are based on the structure 1NA0; for the 42PR sequences, boundaries are based on 4Y6W. (C) Single-helix representation of *Pa* 42PR constructs used in this study, where x signifies the number of internal *Pa* AB units, ranging from one to four. A and B-helices are shown in red and magenta, respectively. Together with connecting turns, these two helices make up a full *Pa* 42PR. N_A and C_B capping helices are shown in grey.

Solution structure of *Pa* 42PRs

To determine the hydrodynamic properties of our *Pa* 42PR constructs, we conducted analytical ultracentrifugation sedimentation velocity (AUC-SV) experiments. We modeled SV data using direct boundary ($\Delta C/\Delta T$) methods (Stafford and Sherwood 2004), as well as c(s) methods (Schuck 2000). The SV data indicate that all constructs populate predominantly (>95%) monomeric species at low (<10μM) concentrations

(Figure 3.3). Higher concentrations result in weak associations; the extent and nature of association differs for each construct (Table 3.1).

$N_AB(AB)AC_B$ and $N_AB(AB)_2AC_B$ SV c(s) distributions are consistent with predominantly monomeric species with higher concentrations displaying small peaks at higher s-values (Figures 3.3A and 3.3C). These higher s-value peaks are shifted from what would be expected for dimeric species. Sedimentation velocity $\Delta C/\Delta T$ curves spanning a wide concentration range were globally fit using a monomer-incompetent trimer model (Figures 3.3B and 3.3D). This model assumes a small proportion of the loading concentration is present as a trimeric species that does not equilibrate with the monomer (Figure S3.1A). Incompetent species have been identified in a number of other AUC studies (Lemaire et al., 2005; Wowor et al., 2011; Xu, 2004). Although the fitted s-values of the incompetent species are consistent with molecular weights corresponding to trimers of each protein, given the low fitted concentrations of these species (less than 3% of the total loading concentration), we cannot rule out the possibility that they result from small amounts of sample impurities.

$N_AB(AB)_3AC_B$ and $N_AB(AB)_4AC_B$ SV c(s) distributions are consistent with predominantly monomeric species with higher concentrations displaying small peaks at s-values consistent with dimeric species (Figures 3.3E and 3.3G). Sedimentation velocity $\Delta C/\Delta T$ curves were

globally fit to a monomer-dimer, incompetent dimer model (Figures 3.3F and 3.3H). This model assumes a rapid and reversible equilibrium between monomer and dimer, and an additional dimeric species that does not equilibrate with the monomer (Figure S3.1B). A summary of hydrodynamic models and parameters used for each construct is shown in Table 1.

Table 3.1. Hydrodynamic properties of *Pa* 42PR constructs

Protein	M_w (kDa)	Model[a]	RMSD[b]	s(A)[c]	$s(A_2)$[c]	$s(A_2')$[c]	$s(A_3')$[c]	K_D (mM)	r (A_n/A) (%)[c,d]
$N_AB(AB)AC_B$	16.5	M + IT	4.50E-03	1.86 (1.85-1.86)	N/A	N/A	3.57 (3.53-3.6)	N/A	1.33 (1.29-1.38)
$N_AB(AB)_2AC_B$	21.1	M + IT	6.56E-03	2.39 (2.37-2.4)	N/A	N/A	4.14 (4.11-4.17)	N/A	2.73 (2.69-2.78)
$N_AB(AB)_3AC_B$	25.8	SA + ID	9.19E-03	2.39 (2.38-2.4)	3.27 (3.24-3.3)	4.08 (4.05-4.12)	N/A	1.2 (1.13-1.27)	1.67 (1.64-1.7)
$N_AB(AB)_4AC_B$	30.4	SA + ID	7.51E-03	2.72 (2.7-2.75)	4.58 (4.47-4.75)	4.88 (4.82-4.94)	N/A	0.27 (0.24-0.32)	2.63 (2.54-2.70)

[a] M, single species monomer; SA, equilibrium self-association; IT, incompetent trimer; ID, incompetent dimer.
[b] Root-mean-squared-deviation of the global fit of the model to the $\Delta C/\Delta T$ data.
[c] 95% confidence intervals calculated from F-statistics analysis (Johnson and Straume, 1994) for fitted parameters are shown in parenthesis.
[d] Ratio of incompetent species to total loading concentration of cell, expressed as a percentage.

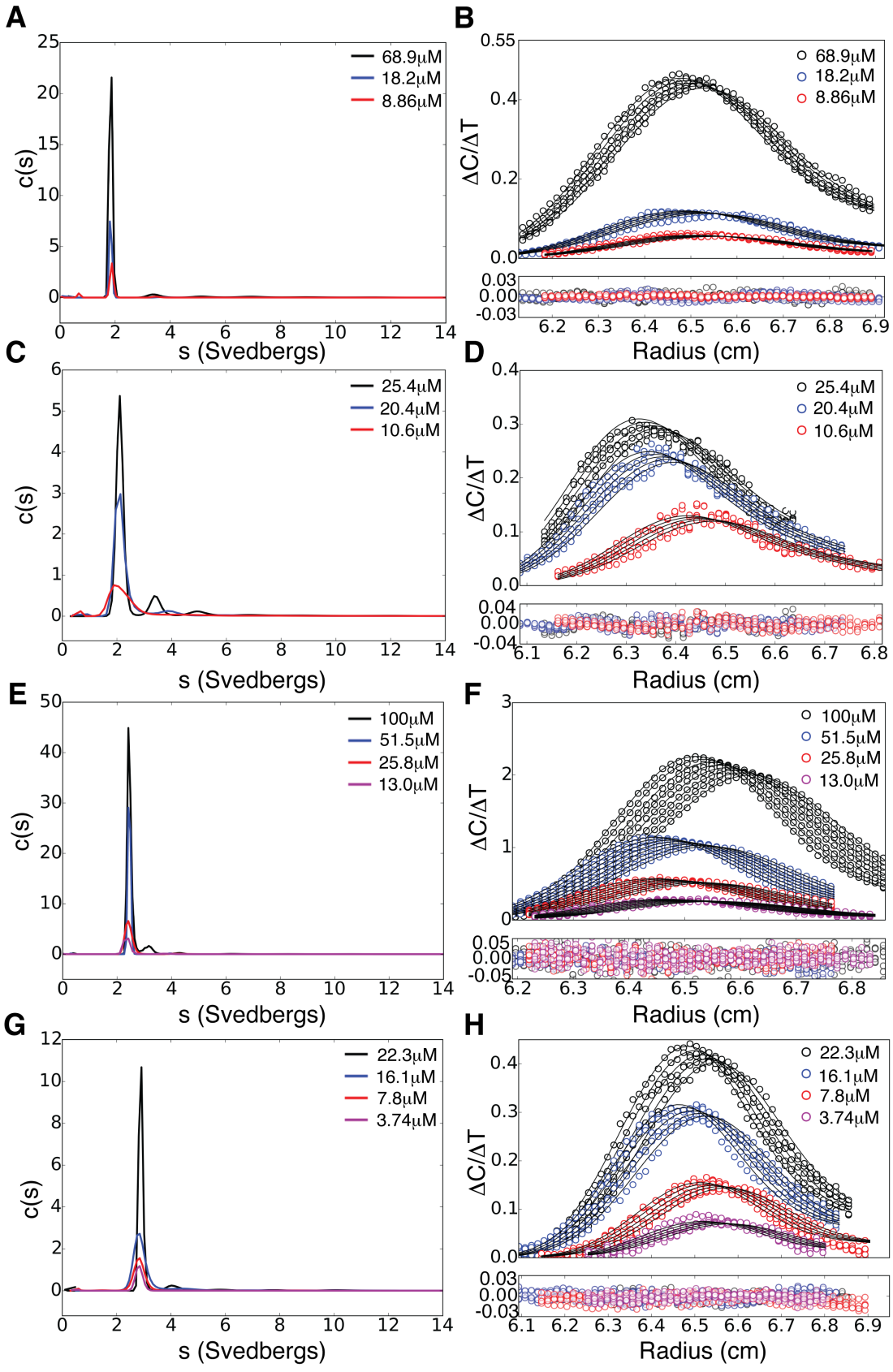

A
B
C
D
E
F
G
H
c(s)
s (Svedbergs)
ΔC/ΔT
Radius (cm)
68.9μM
18.2μM
8.86μM
25.4μM
20.4μM
10.6μM
100μM
51.5μM
25.8μM
13.0μM
22.3μM
16.1μM
7.8μM
3.74μM

Figure 3.3. Sedimentation velocity analytical ultracentrifugation of *Pa* 42PR $N_AB(AB)_xAC_B$ constructs. Continuous c(s) distributions and global $\Delta C/\Delta T$ fits of $N_AB(AB)AC_B$ (panels A,B), $N_AB(AB)_2AC_B$ (panels C,D), $N_AB(AB)_3AC_B$ (panels E,F), and $N_AB(AB)_4AC_B$ (panels G,H), respectively. Lower panels are residuals between $\Delta C/\Delta T$ values and fitted models (Figure S1). For clarity, only a subset of $\Delta C/\Delta T$ values and curves are shown.

Based on sequence elements that match 34PR motifs, we expect the 42PRs derived from *Pa* to adopt an α-helical structure. Consistent with this expectation, far-UV CD spectra of all $N_AB(AB)_xAC_B$ constructs show a high level of α-helical character with well-defined minima at 222 and 208nm (Figure 3.4A). Observed variations in molar residue ellipticity values are likely due to uncertainties in protein concentrations. Each repeat contains only one tyrosine, and no tryptophans (Figure 3.2); thus, extinction coefficients are low. CD spectra for *Pa* 42PR constructs are similar to those of c34PR constructs (Main et al. 2003).

Thermodynamic stability of *Pa* 42PRs

To measure the thermodynamic stability of designed *Pa* 42PRs, we carried out urea-induced equilibrium unfolding of constructs ranging in length from three to six total repeats (Figure 3.4B, closed circles). All transitions are completely reversible. As repeats are added, the transition midpoints increase and become sharper, indicating both increased stability and a high level of cooperativity. These observations are reflected in the fitted free energies and m-values for unfolding (Table 3.2).

To compare stabilities of *Pa* 42PRs to shorter 34PR motifs, we also measured the urea-induced equilibrium unfolding of c34PRs of equivalent numbers of total repeats (Figure 3.4B, open circles). To generate c34PRs with integral numbers of whole repeats, we added c34PR B-helices to the

N-termini of c34PR constructs studied by Regan and coworkers (Main et al., 2003; Kajander et al., 2005; Cortajarena and Regan, 2011). We term these constructs $B(AB)_xS$, where x signifies the number of central c34PRs ranging from two to four, and S is the "solvation helix" designed by Regan and coworkers. The addition of these helices, slightly increases the stability of each construct, and is in good agreement with the intrinsic and interfacial helical coupling energies measured by Regan and coworkers (Kajander et al. 2005), and in excellent agreement with those I measured in Chapter 2.

Although *Pa* 42PRs are larger, they have significantly lower urea midpoints than their c34PR counterparts. *Pa* 42PRs have sharper transitions and increased m-values compared to c34PRs. In contrast with the m-values of c34PRs, which plateau at four repeats, *Pa* 42PR m-values continue to increase through six repeats (Table 3.2). This reflects a higher level of cooperativity in *Pa* 42PRs than in c34PRs. The magnitudes of *Pa* 42PR fitted m-values correlate well with their larger motif size and the expected solvent accessible surface area (SASA) changes for unfolding.

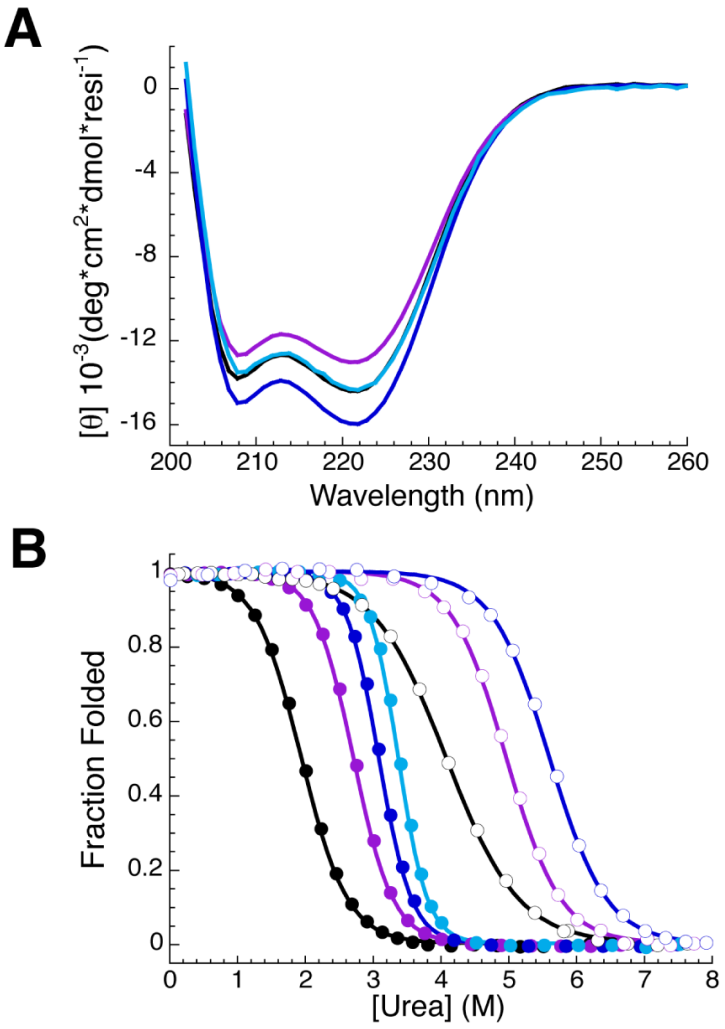


Figure 3.4. Far-UV CD spectroscopy and equilibrium unfolding of 42PR and 34PR constructs. (A) Far-UV CD spectra of *Pa* 42PR constructs $N_AB(AB)AC_B$ (black), $N_AB(AB)_2AC_B$ (purple), $N_AB(AB)_3AC_B$ (blue), and $N_AB(AB)_4AC_B$ (cyan). (B) Normalized equilibrium unfolding transitions of *Pa* 42PR constructs and c34PRs with equivalent numbers of repeats. Closed circles show *Pa* 42PR constructs, and are colored as in (A). Open circles show c34PR constructs: $B(AB)_2S$ (black), $B(AB)_3S$ (purple), and $B(AB)_4S$ (blue). Solid lines result from fitting a two-state unfolding model to the data.

Table 3.2. Thermodynamic parameters of *Pa* 42PR and c34PR constructs

Protein	# of full repeats	$\Delta G°$ (kcal*mol^{-1})	m-value (kcal*mol^{-1}*M^{-1})
$N_AB(AB)AC_B$	3	3.26 ± 0.0038	1.7 ± 0.019
$N_AB(AB)_2AC_B$	4	5.46 ± 0.13	2.05 ± 0.03
$N_AB(AB)_3AC_B$	5	7.63 ± 0.05	2.45 ± 0.023
$N_AB(AB)_4AC_B$	6	9.43 ± 0.22	2.74 ± 0.0073
$B(AB)_2S$	3	4.67 ± 0.18	1.14 ± 0.032
$B(AB)_3S$	4	7.37 ± 0.041	1.5 ± 0.029
$B(AB)_4S$	5	8.13 ± 0.29	1.46 ± 0.073

Parameters are means from three or more independent unfolding transitions fit using a two state unfolding function. Uncertainties represent the standard deviation of the mean. The shaded region highlights *Pa* 42PR constructs and the unshaded region highlights c34PR constructs.

Structure determination of *Pa* 42PR motifs

To determine the atomic structure of tandem 42PR motifs from the *Pa_6_8860* gene product, we crystallized a construct containing five repeats, $N_AB(AB)_3AC_B$. This construct crystallized in space group $P2_12_12$, and contained one molecule in the asymmetric unit, with a calculated solvent fraction of 0.45. Crystals diffracted X-rays past 1.59 Å, with an I/Iσ of ~7 in the highest resolution shell.

Although the data merged with good statistics (Table 3.3), we were unable to solve the structure using molecular replacement with various search models, including single and five repeat arrays of various nPR structures. Anomalous diffraction data collected from selenium-methionine containing protein failed to provide adequate anomalous signal to determine experimental phases, perhaps because the protein contained only two methionine residues located at the N-terminus, outside of the nPR domains.

To obtain a stronger anomalous signal we introduced a single methionine substitution into each of the three central AB repeats. Of four substitution sites tested (L12M, Q17M, N19M, and I35M; numbering is with respect to the position within the 42-residue AB unit), only one substitution site (Q17M) yielded crystals that diffracted to high resolution. Although all protein variants expressed well, N19M and I35M failed to produce crystals, while L12M was largely insoluble. In addition to leucine

being very conserved at this position (Figure 3.1B), the substitution site for L12M is between two helices directly after a sharp turn. Based on hydrophobic packing and sterics, it is unlikely this position can accommodate a residue larger than leucine. Therefore it is plausible L12M destabilizes the helical turn and interface, and results in partial unfolding, leading to the observed insolubility (aggregation).

Collection of single-wavelength anomalous dispersion (SAD) data on Q17M crystals at the Se peak wavelength allowed for determination of experimental phases. These phases produced electron density maps of excellent quality that enabled building and refinement of the Q17M variant structure. Since the Q17M and native protein crystallized in the same space group, the phases and model of Q17M were used to build and refine a model of $N_AB(AB)_3AC_B$ using the native data. Refinement and data collection statistics for both structures are shown in Table 3.3.

The resulting native structure reveals a five repeat right-handed superhelix, with an overall architecture similar to 34PR domains (Figure 3.5). However, each of the A and B helices is approximately one helical turn longer than the 34 residue counterparts. These extensions occur on the N-terminus of the A-helix, and the C-terminus of the B-helix, compared to canonical 34PR helices. The same architecture is maintained across the entire *Pa* 42PR array. Together, an A and B-helix constitute a full 42PR. The average backbone RMSD for all repeats along the array is 1.63

Å. For repeats with identical sequence, the average backbone RMSD is 1.20 Å. Much of this deviation comes from the third central repeat, which differs from the first two by about 1.5 Å. This difference is exemplified by a shorter (6.92 Å) calculated A_i:B_i helical packing distance compared to the first and second central repeats (8.1 and 8.26 Å, respectively), and appears to be a result of slight helical distortions. These helical distortions may be induced by crystal lattice interactions, as there is a large interface (ASA) between symmetry mates that is centered on the third repeat (Figure S3.2). In contrast, the first two central repeats have an RMSD of 0.54 Å. For helices of each type, the backbone RMSD is 1.1 and 0.98 Å, for A- and B-helices, respectively. All possible pairwise repeat alignments are shown in Figure S3.4.

Table 3.3. Data collection and refinement statistics

Crystal	Native	Q17M SeMet
PDB accession code	4Y6W	4Y6C
Wavelength (Å)	1.0375	0.978 (SAD peak)
Refinement resolution range	40.12-1.587 (1.643-1.587)	39.76-1.772 (1.836-1.773)
Space group (*hkl*)	$P2_12_12$	$P2_12_12$
Unit cell dimensions		
a, b, c, (Å)	81.071, 92.341, 30.817	81.103, 91.242, 30.839
α, β, γ (°)	90, 90, 90	90, 90, 90
$R_{sym}/R_{meas}/R_{pim}$	0.112/0.114/0.021 (0.737/0.798/0.153)	0.172/0.175/0.033 (0.348/0.360/0.093)
$CC_{(1/2)}$/CC*	0.999/1 (0.969/0.992)	0.996/0.999 (0.981/0.995)
$\langle I/\sigma I \rangle$	18.6 (7)	17.7 (11.7)
Redundancy	14.2 (14)	14.3 (13.9)
Completeness (%)	99.68 (98.14)	99.92 (99.16)
No. Reflections		
Total	909,077 (85,701)	659,856 (64,472)
Unique	32,123 (3,110)	23,013 (2,236)
R_{work}/R_{free}	0.1778/0.2042 (0.1809/0.2281)	0.1730/0.2135 (0.1885/0.1901)
No. Atoms	1796	1904
Protein	1681	1725
Water/solvent	115	169
RMS deviations		
Bond lengths (Å)	0.005	0.006
Bond angles (Å)	0.95	0.92
Ramachandran analysis		
Most favored (%)	99	99
Allowed (%)	1	1

Values enclosed in parenthesis represent the highest resolution shell.

$R_{sym} = \Sigma_{hkl} \mid I(hkl) - \langle I(hkl) \rangle \mid / \Sigma_{hkl} I(hkl)$

$R_{meas} = \Sigma_{hkl} \sqrt{(n/(n-1))} \mid I(hkl) - \langle I(hkl) \rangle \mid / \Sigma_{hkl} I(hkl)$

$R_{meas} = \Sigma_{hkl} \sqrt{(1/(n-1))} \mid I(hkl) - \langle I(hkl) \rangle \mid / \Sigma_{hkl} I(hkl)$

$R_{work} = \Sigma_{hkl} \mid F_{obs} - F_{calc} \mid / \Sigma_{hkl} F_{obs}$; R_{free} = test set 6.23% (Native) and 5.14% (Q17M)

$CC^* = \sqrt{(2CC_{(1/2)})/(1 + CC_{(1/2)})}$ (Karplus and Dieterichs, 2012)

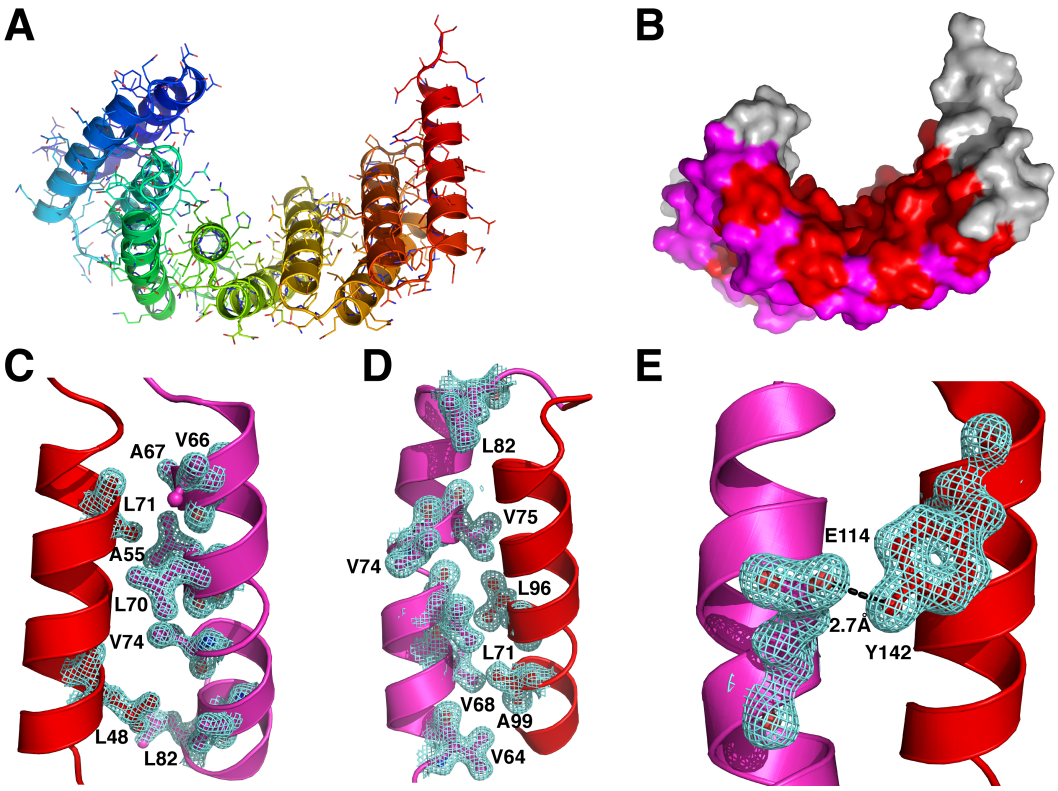


Figure 3.5. Crystal Structure of *Pa* 42PR $N_AB(AB)_3AC_B$. (A) Cartoon representation of the $N_AB(AB)_3AC_B$ crystal structure 4Y6W, colored from N-terminus (blue) to C-terminus (red). Crystallographic waters are omitted for clarity. (B) Surface representation of $N_AB(AB)_3AC_B$ molecule. View is same as in (A), with coloring scheme as in Figure 3.2C. Each 42-residue repeat includes an A and B helix. (C-E) Representative electron density ($2F_O$-F_C, contoured at one sigma) of interactions along the repeat array. (C) and (D) Hydrophobic residues in an intra-repeat (A_i:B_i) and inter-repeat (B_i:A_{i+1}) helical interface, respectively. (E) One of four conserved Tyr $O\eta H$--$O\varepsilon C$ Glu hydrogen bonds present within each of the inter-repeat helical interfaces.

3.4 Discussion

A nomenclature system for variable-length TPR-like motifs

The TPR sequence was originally identified as a 34-residue motif in *Saccharomyces cerevisiae* cell-cycle regulation machinery (Sikorski et al., 1990). The structure of the 34 residue TPR motif, revealed by Das et al. (Das et al., 1998), consists of two anti-parallel α-helices (A and B). A large number of 34 residue TPRs have since been added to this family, and conform closely in sequence and structural features. However, as databases have grown, TPR-like sequences have appeared that differ from the canonical length. The nPR nomenclature introduced here (where n represents the number of residues in the repeating unit) captures this length variation.

Sequence features of the 42PR motif

We identified a new class of nPR sequences, which we term 42PRs. The repeating unit consists of consecutive 42 residue segments, which share sequence characteristics with canonical 34PRs (Figure 3.1). The main differences between the two sequence classes appear to be N and C-terminal extensions of the 34PR-defined A- and B-helices, respectively, in 42PRs. The 42PR HMM shows a high degree of conservation near the N terminus. This conservation may reflect the sequence characteristics defining a longer A-helix in 42PRs. The C-

terminus of the 42PR HMM shows less conservation, indicating the rules defining the extension of the B-helix in 42PRs may be less strict.

Conserved positions in both 42PRs and 34PRs include small and hydrophobic residues at helix interfaces and in turn regions (Figure 3.1). Based on these structurally restrictive environments, it is likely that the conserved nPR residues are responsible for defining the fold. These residues may act as staples, around which helical extensions can be accommodated. In this fashion, the structural registry and packing of conserved nPR residues between helices is maintained.

Structural features of nPR motifs

Due to the repetitive architecture of nPR proteins, their structures can be defined by a small number of repeating parameters: helix crossing angles, distances, and contacts. These parameters, and the tertiary structures they define, are important for function and for stability. Helix crossing angles determine the extent to which nPR arrays form a concave binding surface for target peptides and proteins (Cortajarena and Regan, 2006; Cortajarena et al., 2010; Zhu et al., 2012). The extent to which contacts are formed between nPR helices (A_i:B_i, B_i:A_{i+1}, and A_i:A_{i+1}) likely contributes to cooperativity in folding. By comparing the structural and energetic features of consensus, naturally occurring, and highly repetitive 42- and 34PRs, we can determine which structural parameters are general

to nPRs, to 42- versus 34PRs, and which are modulated by local sequence variation within families.

Aside from differences in helix lengths, many of the structural features of 42PRs are similar to those of 34PRs. Although there are small deviations from repeat to repeat, average crossing angles for the A_iB_i, B_i:A_{i+1}, and A_i:A_{i+1} helices are similar (Table 3.4). Likewise, the helical distances and solvent accessible surface area (SASA) burial between helices are similar across nPR families (Table 3.4).

Table 3.4. Helix-helix interfaces in representative nPRs

	4Y6W (*Pa* 42PR)	3CEQ (42PR)	1NA0 (c34PR)	1ELW (34PR)
Average A_i:B_i helix crossing angle (°)	159 ± 6.8	159.3 ± 5.3	162.2 ± 0.8	169.6 ± 6.8
Average B_i:A_{i+1} helix crossing angle (°)	155 ± 2.3	165.7 ± 6.2	154 ± 1.4	156.2 ± 6.1
Average A_i:A_{i+1} helix crossing angle (°)	20.5 ± 7.4	24.8 ± 1.4	30.6 ± 1.4	23.4 ± 5.2
Average A_i:B_i packing distance (Å)	7.54 ± 0.62	7.37 ± 1.13	7.3 ± 0.01	6.92 ± 1.2
Average B_i:A_{i+1} packing distance (Å)	8.5 ± 0.03	7.91 ± 2.05	8.78 ± 0.16	8.69 ± 0.78
Average A_i:A_{i+1} packing distance (Å)	10.9 ± 0.97	12.4 ± 0.44	11 ± 0.22	9.93 ± 1.7
Average SASA burial in A_i:B_i (Å^2)	1308 ± 102	1422 ± 169	1363 ± 35	1260 ± 67
Average SASA burial in B_i:A_{i+1} (Å^2)	1571± 78	1400 ± 131	1291 ± 18	1252 ± 123
Average pairwise sequence ID between repeats	97%	48%	96%	22%
Total number of repeats	5	4.5	3.5	3.5

Uncertainties represent standard errors on the mean. The shaded region highlights 42PRs; the unshaded region highlights 34PRs.

In contrast, there are several notable differences in contacts within and between 42- and 34PRs (Figure 3.6). In addition to contacts within helices (main diagonals), nPR structures show a characteristic contact pattern, consisting mainly of contacts between successive helices (A_i:B_i and B_i:A_{i+1}; anti-diagonal features). Other contacts include those between successive A-helices (A_i:A_{i+1}; off-diagonal features), but are less frequent in 42PR structures than in 34PR structures. These A_i:A_{i+1} contacts connect the regularly spaced, anti-diagonal contacts (Figure 3.6).

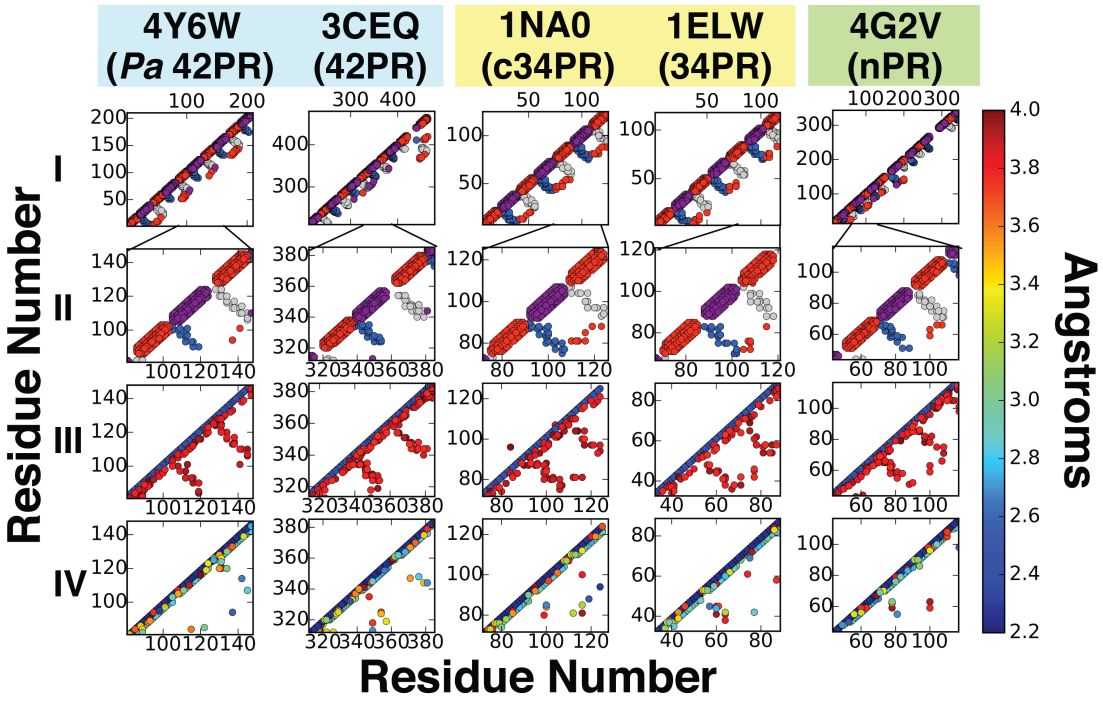


Figure 3.6. Contact maps of 42 and 34 residue nPRs. Contacts are defined as atom pairs from different residues that are within 2.2-4 angstroms (see color bar). Points above the main diagonal represent backbone contacts. Points below the main diagonal represent backbone-side chain or side chain-side chain contacts. Protein structures are displayed in each column (cyan, 42PRs; yellow, 34PRs; green, both 42PR and 34PR repeats). Row I displays helical contacts over the entire structure. Rows II, III, and IV expand over the indicated residue range. Rows I and II display contacts within A-helices (red), B-helices (magenta), between A_i and B_i helices (blue), and between B_i and A_{i+1} helices (grey). Contacts between A_i and A_{i+1} helices appear as red, off-diagonal points in Rows I and II. Rows III and IV show all hydrophobic and polar contacts, respectively, and are color coded with respect to distance.

The packing features describing A_i:A_{i+1} helices can be visualized by alignment of $A_iB_iA_{i+1}$ units from each motif with respect to A_iB_i, and are summarized in Table 3.4. Although the helix geometries within nPR families are similar, 42PRs have fewer A_i:A_{i+1} contacts than 34PRs. For 3CEQ and 1ELW (naturally occurring 42- and 34PRs, respectively), the differences in A_i:A_{i+1} contacts can be explained by local helix geometry (longer A_i:A_{i+1} distances in 3CEQ). Local helix geometry in *Pa* 42PRs also affects the number of A_i:A_{i+1} contacts, as the N_A-helix kink results in more A_i:A_{i+1} contacts in the N-terminal region of the contact plot, relative to the internal repeat region (Figure 6, row I). Despite these local effects, *Pa* 42PRs and c34PRs have overall similar A_i:A_{i+1} helix distances. This suggests that sequence specific information also influences the extent of A_i:A_{i+1} interaction.

To further analyze the types of interactions in the A_i:B_i, B_i:A_{i+1}, and A_i:A_{i+1} interfaces, we sorted polar and nonpolar interactions into separate contact maps (Figure 3.6, rows III and IV). Sorted contact plots reveal that packing within all interfaces, both within (A_i:B_i) and between (B_i:A_{i+1}) repeats, is predominantly hydrophobic, although these hydrophobic contacts (Figures 3.5C, 3.5D, and Figure 3.6, row III) occur at greater distances than the polar contacts (Figure 3.6, row IV). A few polar contacts are also present among side chains forming the B_i:A_{i+1} interfaces of *Pa* 42PRs, including a conserved Tyr OηH--$^-$OεC Glu hydrogen bond on the

convex side of the *Pa* 42PR superhelix, between adjacent 42PR motifs (Figures 3.1B, 3.5E). Other repetitive polar interactions include a His-Ser-Gln hydrogen bond network connecting successive $A_iB_iA_{i+1}$ helices (Figure S3.4) and a His-Ser A-helix N-terminal capping motif (Figure S3.5).

Interestingly, the 42PRs of the human kinesin light chain (3CEQ) contain a region with B_i:B_{i+1} helical contacts (Figure 3.6, Row I), which are not seen in 34PRs. This results from a slight kink in one B-helix, allowing for enhanced hydrophobic packing, along with other contacts between polar residues on the convex face of the superhelix. These sequence and length-specific structural variations highlight the structural malleability of the nPR motif. In naturally occurring repeat proteins, sequence variation can locally tune structural features. A consensus design approach applied to 42PRs would be expected to reveal representative interactions across all 42PRs.

Possible functions of the 42PR family genes and the implications of identical repeats

Although the function of *Pa_6_8860* is unclear, many homologous sequences share a common architecture containing different N-terminal domains, flanked by nPRs and other repeat protein types near the C-terminus (van der Biezen and Jones, 1998). The annotated functions of these proteins range from apoptosis and cell death regulation to plant

resistance. The N-terminal portion of *Pa_6_8860* is predicted to contain a partial NB-ARC domain, although it lacks some of the key residues involved in binding ATP (Yan et al., 2005). The 42PRs of *Pa_6_8860* show the greatest sequence identity to the human kinesin light chain (KLC) nPRs, and there is evidence that many KLC domains contain nPRs (Pernigo et al., 2013; Fischer et al., 2012; Zhu et al., 2012; Gindhart and Goldstein, 1996); a subset of these have been shown to bind to cargo. Thus, it is plausible that *Pa_6_8860* functions as a kinesin light chain, and the *Pa* 42PRs may be involved in cargo binding in microtuble-based vesicular transport.

In the crystal lattice, there is an extensive interface between symmetry mates, involving the concave face of the 42PR array (Figure S3.2). This interface buries ~9360 $Å^2$ of total SASA. It is possible this dimer reflects the one characterized by AUC-SV, which has a fitted K_D of 1.2mM. Interestingly, the addition of a single repeat to this protein results in a ~4.5-fold tighter dimerization K_D. It is therefore possible the 15 tandem *Pa* 42PRs in *Pa_6_8860* have the potential to form even tighter interactions. If this dimerization surface overlaps the cargo binding surface, dimerization and cargo binding would likely be competitive, and thus, cargo binding may dissociate kinesin light-chains. This explanation could also explain cargo delivery, as the competition would work in

reverse in regions of lower concentration of cargo relative to *Pa* 42PRs after transport.

Due to the high internal sequence identity and abnormally large number (15) of *Pa* 42PRs in *Pa_6_8860*, the 42PR domain is expected to have multiple identical binding sites. These identical sites would have the potential to display an avidity effect for polyvalent targets (with direct sequence and/or structural repetition). An example of direct tandem repeats binding to a repetitive target is the TALE repeats of plant pathogenic bacteria, which bind to duplex DNA (Boch et al., 2009; Deng et al., 2012; Mak et al., 2012).

Due to the high internal sequence identity of the *Pa* 42PRs in *Pa_6_8860*, it is possible the resulting 42PR domain has multiple identical binding sites. These identical sites would have the potential to display an avidity effect for polyvalent targets (with direct sequence and/or structural repetition). An example of direct tandem repeats binding to a repetitive target is the TALE repeats of plant pathogenic bacteria, which bind to duplex DNA (Boch et al., 2009; Deng et al., 2012; Mak et al., 2012). Such an avidity effect would be further enhanced if there were attractive interactions among ligands. In addition, energetic coupling among sites has the potential to display cooperative effects.

Folding of *Pa* 42PRs

The cooperative folding of the *Pa* 42PR arrays in this study is striking. As repeats are added, both stabilities and m-values increase. This phenomenon is characteristic of other linear repeat proteins, especially ankyrin repeats, where energetic coupling leads to highly cooperative folding (Aksel et al., 2011; Wetzel et al., 2008). In contrast, the unfolding transitions of c34PRs plateau at four repeats, consistent with a high level of partially folded states (Kajander et al., 2005; Cortajarena and Regan 2011) and decreased cooperativity compared to ankyrin repeats and the *Pa* 42PRs presented here.

The magnitude of the unfolding cooperativity can be directly measured by applying a one-dimensional Ising (nearest-neighbor) model (Aksel and Barrick, 2009; Kajander et al., 2005; Mello and Barrick, 2004; Wetzel et al., 2008). This approach is usually limited to designed repeat proteins, as repeats must have identical sequences. This requirement prevents the application of these models to natural repeat protein folding, where sequence identity between repeats is typically very low (~25%). The *Pa* 42PRs we describe in this study are ideal candidates for analysis using nearest-neighbor models. Such analysis will provide an understanding of how natural nPR proteins distribute their energy to fold cooperatively.

3.5 Experimental Procedures

Subcloning, protein expression, and purification

DNA sequences encoding AB, N_AB, and AC_B repeats (Figure 3.2B) were cloned by annealing complementary, codon optimized oligonucleotides. Annealed single-repeat cassettes were ligated directly into NdeI and BglII digested pET-15b (Novagen, Madison, WI). BamHI sites were included in the AB and AC_B cassettes to allow for ligation as previously described (Aksel et al, 2012). Single-site substitutions (L12M, Q17M, N19M, and I35M) were introduced using Quikchange (Stratagene, La Jolla, CA) on individual AB cassettes. c34PR constructs were created using the same approach. DNA sequences for all constructs in this study are listed in Table S3.1.

Pa 42PR constructs were expressed in *Escherichia coli* Rosetta R2* (DE3) cells. One liter cultures were grown in terrific broth to an OD_{600} of 0.8, induced by adding IPTG to 200 μM, and incubated overnight at 20°C. Bacteria were pelleted, and lysed in 50 mL 25 mM Tris-HCl, 350 mM NaCl, 25 mM imidazole, 10 mM $MgCl_2$ pH 8, 1 mg DNase, and tagged proteins were purified from the supernatant via Ni-NTA chromatography. Purified proteins were dialyzed extensively into 25 mM Tris-HCl, 350 mM NaCl pH 8, concentrated using an Amicon stirred cell concentrator (EMD Millipore, USA), and flash frozen at -80°C. Protein concentrations were determined as described by Edelhoch (Edelhoch, 1967).

To express selenium-methionine–substituted proteins, cells were pelleted at an OD_{600} of 0.8, and were resuspended in M9 medium containing 100 mg/L selenium-methionine (Acros Organics, USA), 500 mg/L lysine, phenylalanine, and threonine, 250 mg/mL isoleucine, leucine, and valine (inhibitory amino acids for methionine biosynthesis), and 200 μM IPTG. During purification and analysis, 5 mM TCEP was included to selenium-methionine-substituted protein samples to ensure reduction of selenium. Complete selenium-methionine incorporation was confirmed using mass spectrometry.

Circular dichroism spectroscopy

CD measurements were conducted using an Aviv Model 400 CD Spectropolarimeter (Lakewood, NJ). CD samples contained 25 mM Tris-HCl, 350 mM NaCl, pH 8. Far-UV CD spectra were recorded at 25°C using a 0.1 cm path-length quartz cuvette (Starna Cells Inc., Atascadero, CA) at protein concentrations ranging from 15-25 μM. Spectra were obtained by signal averaging every 1 nm for 30 s. Buffer spectra in the same cuvette were subtracted prior to analysis.

Urea-induced equilibrium unfolding transitions

Equilibrium unfolding transitions were obtained by monitoring CD at 222 nm (*Pa* 42PRs) and 220 nm (c34PRs) in a 1 cm path-length quartz

cuvette. High purity urea (Amresco, Solon, OH) was deionized by stirring with mixed-bed resin (Bio-Rad, Hercules, CA) as previously described in (Street et al., 2008). Urea concentration was determined by refractometry (Pace 1986). Titrations were performed at protein concentrations ranging from 1.0-2.5 μM using a computer-controlled Microlab syringe titrator (Hamilton, Reno, NV). At each urea concentration, protein samples were equilibrated for 5–7 min, and the CD signal was averaged for 30 s. Two-state analysis of equilibrium unfolding transitions were carried out as described (Street et al., 2008).

Analytical Ultracentrifugation

Analytical ultracentrifugation sedimentation velocity (AUC-SV) experiments were performed using a ProteomeLab-equipped Beckman XL-I analytical ultracentrifuge. Prior to AUC experiments, all proteins were extensively dialyzed into 25 mM Tris-HCl, 350 mM NaCl, pH 8. Protein samples were prepared to span a wide concentration range (~5-100 μM) by dilution with the dialysate.

AUC-SV cells were assembled using SedVel60K 1.2 mm meniscus-matching centerpieces (SpinAnalytical) and sapphire windows. All other cell components were purchased from Beckman Coulter. Upon sample and reference (dialysate) loading, centerpieces were aligned in a An-60Ti rotor, and menisci were matched according to (Allgood and Barrick, 2011).

After remixing, the rotor was thermally equilibrated under vacuum at 25 °C for at least 90 minutes. SV experiments were run for approximately 8 hours at 45-50krpm.

Protein Crystallization and Data Collection

Crystals of native $N_AB(AB)_3AC_B$ were grown at room temperature (~22 °C) by hanging-drop vapor diffusion. Protein solution (20 mg/mL) was mixed in either a 2:1 or 1:1 ratio with reservoir solution containing 0.1 M MES (pH 6.5) and 25-30% PEG 4K. Crystals appeared after approximately 3-7 days. Crystals were cryoprotected by transfer into a solution consisting of 0.1 M MES (pH 6.5), 35% PEG 4K, and 5-10% ethylene glycol, and then flash frozen in liquid nitrogen. The selinium-methionine-substituted Q17M variant gave rise to morphologically similar crystals under these conditions, with the addition of 5 mM TCEP.

Native and selenium derivative data sets were collected at the National Synchrotron Light Source (NSLS) beamlines X-25 and X-29 (Brookhaven National Laboratory, Brookhaven, NY)) and processed with HKL2000 (Otwinowski and Minor, 1997). Crystals belong to space group $P2_12_12$ and contain one molecule per asymmetric unit.

Structure Determination and Analysis

Selenium positions were determined by SAD in ShelXC/D (Sheldrick, 2010; Schneider and Sheldrick, 2002) through HKL2MAP (Pape and Schneider 2004). Phases were calculated in SOLVE and improved by density modification in RESOLVE (Terwilliger 2003). Iterative rounds of building and refinement were performed using COOT (Emsley and Cowtan 2004), Phenix (Adams et al. 2010; Afonine et al. 2012; Afonine et al. 2009; Afonine et al. 2013; Headd et al. 2012), and Refmac (Murshudov et al., 1997). The final model was validated with the program Molprobity (Chen et al., 2010). The native structure, which crystallized in the same unit cell as the SeMet derivative, was built from the refined Q17M structure as a starting model using Phaser (McCoy et al., 2007). Molecular images were generated using PyMOL Version 1.5.0.4 (Schrödinger, LLC). Helix crossing angles were calculated using helix_angles.py (R.L. Campbell, Queens University). Solvent accessible surface area calculations were performed using MSMS (Sanner et al., 1996). Structural alignments were performed using LSQMAN (Kleywegt, 1996).

Acknowledgements: We would like to thank Drs Annie Héroux and Howard Robinson of the Macromolecular Crystallography Research Resource (PXRR) at the National Synchrotron Light Source, Brookhaven,

NY for their technical assistance with data collection strategy and preliminary analysis. We also thank Dr. Phil Mortimer of the JHU mass spectrometry facility and Dr. Michael Love of the JHMI X-ray facility. This work was supported by NIH grant R01 GM068462 to D.B., J.D.M. was supported by NIH training grant T32-GM008403, J.M.K. was supported by NIH grant R01 GM099231 to Daniel J. Leahy, and G.D.B. was supported by NIH grant R01 GM084192.

3.6 Supplemental information

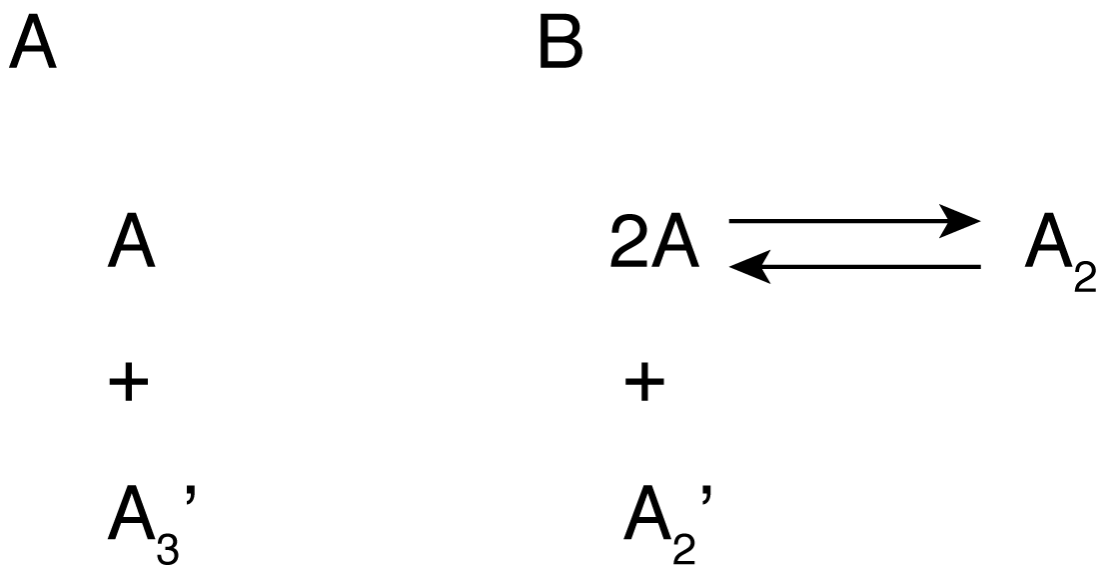


Figure S3.1. Schematic representation of *Pa* 42PR hydrodynamic models.

(A) M + IT model (Table 3.1) represents a mixture of a non-equilibrating monomer (A) and trimer (A_3'). (B) SA + ID model (Table 3.1) represents a reversible monomer-dimer equilibrium and a non-equilibrating dimer (A_2').

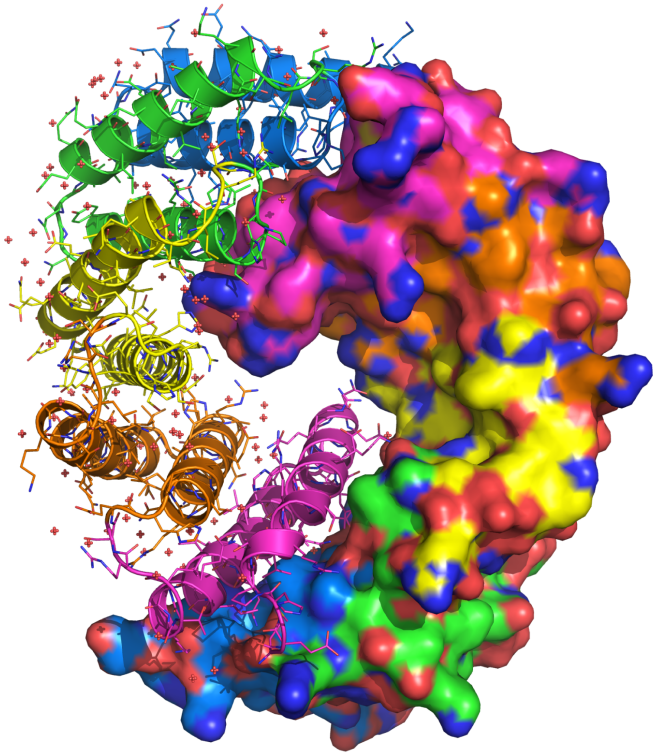


Figure S3.2. Crystallographic 4Y6W dimer.

Observed crystallographic dimer of 4Y6W with ribbon and surface representation. Repeats are colored as follows: N_AB (magenta), AB_1 (orange), AB_2 (yellow), AB_3 (green), and AC_B (blue). Red dots represent crystallographic waters.

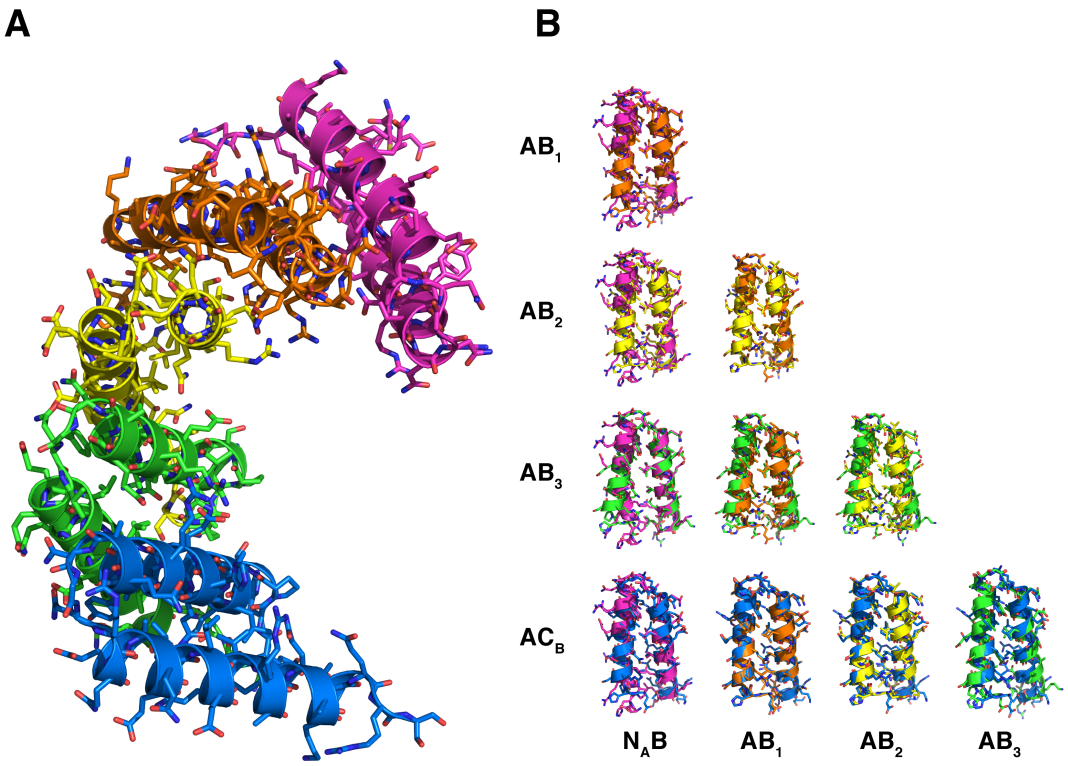


Figure S3.3. Structural alignments of *Pa* 42PRs.

(A) Crystal structure of 4Y6W with N_AB (magenta), AB_1 (orange), AB_2 (yellow), AB_3 (green), and AC_B (blue) repeat coloring. (B) Structural alignments of possible pairwise repeat combinations, colored as in (A).

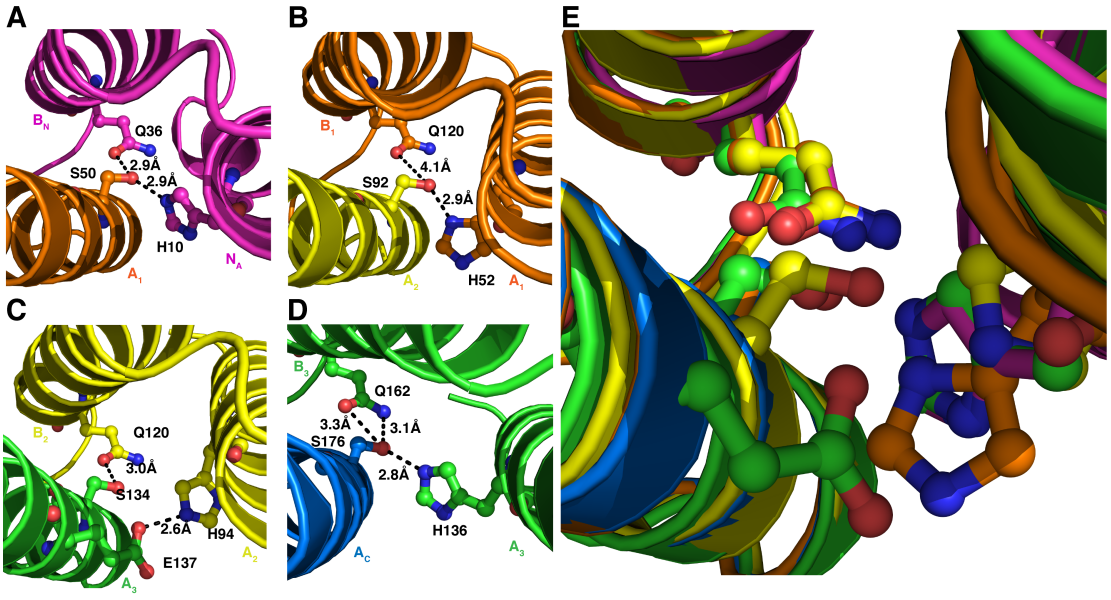


Figure S3.4. His-Ser-Gln H-bond network in *Pa* 42PRs.

(A-D) Hydrogen bond networks between successive A-B-A helices. (A) N_AB_N:A_1 interface, (B) AB_1:A_2 interface, (C) AB_2:A_3 interface, and (D) AB_3:A_C interface. In (A), (B), and (D), hydrogen bond networks center on an A-helix serine, which appears to form a bifurcated hydrogen bond with a histidine and a glutamine from the A and B-helix, respectively, of the previous repeat. In (C), the H94-S134 H-bond is stolen by E137 acting as the acceptor. (E) All His-Ser-Gln H-bonds in $N_AB(AB)_3AC_B$. Helices are colored as in Figure S3.3A.

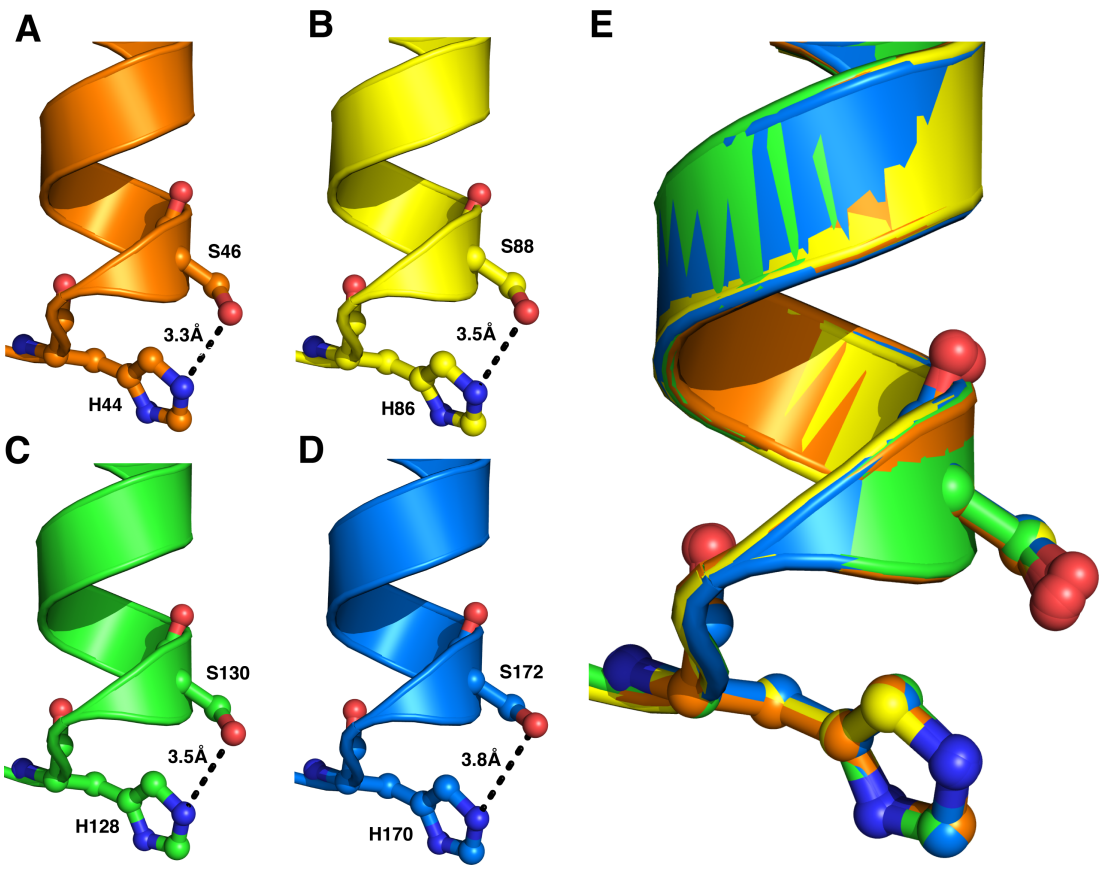


Figure S3.5. His-Ser helix capping H-bond in *Pa* 42PRs.

(A-D) Histidine-Serine hydrogen bond on N terminus of A-helices on concave surface of the $N_AB(AB)_3AC_B$ superhelix. (E) Alignment of all His-Ser N-terminal helix capping H-bonds in $N_AB(AB)_3AC_B$. Helices are colored as in Figure S3.3A.

Table S3.1 *Pa* 42PR and c34PR DNA sequences

Protein name	DNA sequence
$N_AB(AB)AC_B$	atgggccatatggaacatccgagccgcctgcgcagccagcatgaactggcgcgccgctatcagcagaacggccaggtgcag gaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccgcctggcgagc cagcatgaactggcgcgcgcgtatcaggcgaacggccaggtgcaggaagcggtggaactgctggaacaggtggtggcgatt caggcgaaaaccctgagatccgaacatccgagccgcctggcgagccagcatgaactggcgcgcgcgtatcaggcgaacgg ccagcgccaggaagcgcaggaactgctggaacaggtgcgcgcgattcaggcgaaaacccagagatctctggtgccgcgcg gcagcggcagcagccatcatcatcatcatcat
$N_AB(AB)_2AC_B$	atgggccatatggaacatccgagccgcctgcgcagccagcatgaactggcgcgccgctatcagcagaacggccaggtgcag gaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccgcctggcgagc cagcatgaactggcgcgcgcgtatcaggcgaacggccaggtgcaggaagcggtggaactgctggaacaggtggtggcgatt caggcgaaaaccctgagatccgaacatccgagccgcctggcgagccagcatgaactggcgcgcgcgtatcaggcgaacgg ccaggtgcaggaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccg cctggcgagccagcatgaactggcgcgcgcgtatcaggcgaacggccagcgccaggaagcgcaggaactgctggaacag gtgcgcgcgattcaggcgaaaacccagagatctctggtgccgcgcggcagcggcagcagccatcatcatcatcatcat
$N_AB(AB)_3AC_B$	atgggccatatggaacatccgagccgcctgcgcagccagcatgaactggcgcgccgctatcagcagaacggccaggtgcag gaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccgcctggcgagc cagcatgaactggcgcgcgcgtatcaggcgaacggccaggtgcaggaagcggtggaactgctggaacaggtggtggcgatt caggcgaaaaccctgagatccgaacatccgagccgcctggcgagccagcatgaactggcgcgcgcgtatcaggcgaacgg ccaggtgcaggaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccg cctggcgagccagcatgaactggcgcgcgcgtatcaggcgaacggccaggtgcaggaagcggtggaactgctggaacagg tggtggcgattcaggcgaaaaccctgagatccgaacatccgagccgcctggcgagccagcatgaactggcgcgcgcgtatca ggcgaacggccagcgccaggaagcgcaggaactgctggaacaggtgcgcgcgattcaggcgaaaacccagagatctctg gtgccgcgcggcagcggcagcagccatcatcatcatcatcat
$N_AB(AB)_4AC_B$	atgggccatatggaacatccgagccgcctgcgcagccagcatgaactggcgcgccgctatcagcagaacggccaggtgcag gaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccgcctggcgagc cagcatgaactggcgcgcgcgtatcaggcgaacggccaggtgcaggaagcggtggaactgctggaacaggtggtggcgatt caggcgaaaaccctgagatccgaacatccgagccgcctggcgagccagcatgaactggcgcgcgcgtatcaggcgaacgg ccaggtgcaggaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccg cctggcgagccagcatgaactggcgcgcgcgtatcaggcgaacggccaggtgcaggaagcggtggaactgctggaacagg tggtggcgattcaggcgaaaaccctgagatccgaacatccgagccgcctggcgagccagcatgaactggcgcgcgcgtatca ggcgaacggccaggtgcaggaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaac atccgagccgcctggcgagccagcatgaactggcgcgcgcgtatcaggcgaacggccagcgccaggaagcgcaggaact gctggaacaggtgcgcgcgattcaggcgaaaacccagagatctctggtgccgcgcggcagcggcagcagccatcatcatcat catcat
Q17M	atgggccatatggaacatccgagccgcctgcgcagccagcatgaactggcgcgccgctatcagcagaacggccaggtgcag gaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccgcctggcgagc cagcatgaactggcgcgcgcgtatatggcgaacggccaggtgcaggaagcggtggaactgctggaacaggtggtggcgattc aggcgaaaaccctgagatccgaacatccgagccgcctggcgagccagcatgaactggcgcgcgcgtatatggcgaacggc caggtgcaggaagcggtggaactgctggaacaggtggtggcgattcaggcgaaaaccctgagatccgaacatccgagccgc ctggcgagccagcatgaactggcgcgcgcgtatatggcgaacggccaggtgcaggaagcggtggaactgctggaacaggtg gtggcgattcaggcgaaaaccctgagatccgaacatccgagccgcctggcgagccagcatgaactggcgcgcgcgtatcag gcgaacggccagcgccaggaagcgcaggaactgctggaacaggtgcgcgcgattcaggcgaaaacccagagatctctggt gccgcgcggcagcggcagcagccatcatcatcatcatcat
$B(AB)_2S$	catatgtatgatgaagcgattgaatattaccagaaagcgctggaactggacccgagatccgcggaagcctggtataacctgggt aacgcgtattacaaacagggcgattacgacgaagcgatcgaatattaccagaaagcgctggaactggacccgagatccgcg gaagcctggtataacctgggtaacgcgtattacaaacagggcgattacgacgaagcgatcgaatattaccagaaagcgctgg aactggacccgagatccgctgaagctaaacaaaatcttggtaatgctaaacaaaaacaaggttaagatcc
$B(AB)_3S$	catatgtatgatgaagcgattgaatattaccagaaagcgctggaactggacccgagatccgcggaagcctggtataacctgggt aacgcgtattacaaacagggcgattacgacgaagcgatcgaatattaccagaaagcgctggaactggacccgagatccgcg gaagcctggtataacctgggtaacgcgtattacaaacagggcgattacgacgaagcgatcgaatattaccagaaagcgctgg aactggacccgagatccgcggaagcctggtataacctgggtaacgcgtattacaaacagggcgattacgacgaagcgatcga atattaccagaaagcgctggaactggacccgagatccgctgaagctaaacaaaatcttggtaatgctaaacaaaaacaaggtt aagatcc
$B(AB)_4S$	atgggccatcatcatcatcatcatcatcacagcagcggccatatcgaaggtcgtcatatgtatgatgaagcgattgaatattac cagaaagcgctggaactggacccgagatccgcggaagcctggtataacctgggtaacgcgtattacaaacagggcgattacg acgaagcgatcgaatattaccagaaagcgctggaactggacccgagatccgcggaagcctggtataacctgggtaacgcgta ttacaaacagggcgattacgacgaagcgatcgaatattaccagaaagcgctggaactggacccgagatccgcggaagcctg gtataacctgggtaacgcgtattacaaacagggcgattacgacgaagcgatcgaatattaccagaaagcgctggaactggacc cgagatccgcggaagcctggtataacctgggtaacgcgtattacaaacagggcgattacgacgaagcgatcgaatattaccag aaagcgctggaactggacccgagatccgctgaagctaaacaaaatcttggtaatgctaaacaaaaacaaggttaagatcc

3.7 References

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D Biological Crystallography *66*, 213–221.

Afonine, P.V., Grosse-Kunstleve, R.W., Urzhumtsev, A., and Adams, P.D. (2009). Automatic multiple-zone rigid-body refinement with a large convergence radius. Journal of Applied Crystallography *42*, 607–615.

Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T.C., Urzhumtsev, A., Zwart, P.H., and Adams, P.D. (2012). Towards automated crystallographic structure refinement with *phenix.refine*. Acta Crystallographica Section D Biological Crystallography *68*, 352–367.

Afonine, P.V., Grosse-Kunstleve, R.W., Adams, P.D., and Urzhumtsev, A. (2013). Bulk-solvent and overall scaling revisited: faster calculations, improved results. Acta Crystallographica Section D Biological Crystallography *69*, 625–634.

Aksel, T., and Barrick, D. (2009). Chapter 4 Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models. In Methods in Enzymology, (Elsevier), pp. 95–125.

Aksel, T., Majumdar, A., and Barrick, D. (2011). The Contribution of Entropy, Enthalpy, and Hydrophobic Desolvation to Cooperativity in Repeat-Protein Folding. Structure *19*, 349–360.

Allgood, A.G., and Barrick, D. (2011). Mapping the Deltex-Binding Surface on the Notch Ankyrin Domain Using Analytical Ultracentrifugation. Journal of Molecular Biology *414*, 243–259.

van der Biezen, E.A., and Jones, J.D. (1998). The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. Current Biology *8*, R226–R228.

Binz, H.K., Stumpp, M.T., Forrer, P., Amstutz, P., and Plückthun, A. (2003). Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. Journal of Molecular Biology *332*, 489–503.

Blatch G.L., Lässle M. (1999). The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. BioEssays 21, 932-939.

Boch J., Scholze H., Schornack S., Landgraf A., Hahn S., Kay S., Lahaye T., Nickstadt A., Bonas U. (2009). Breaking the code of DNA binding specificity of TAL-Type III effectors. Science 326, 1509-1512.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). *MolProbity*: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D Biological Crystallography *66*, 12–21.

Cortajarena, A.L., and Regan, L. (2006). Ligand binding by TPR domains. Protein Science *15*, 1193–1198.

Cortajarena, A.L., and Regan, L. (2011). Calorimetric study of a series of designed repeat proteins: Modular structure and modular folding. Protein Science *20*, 336–340.

Cortajarena, A.L., Wang, J., and Regan, L. (2010). Crystal structure of a designed tetratricopeptide repeat module in complex with its peptide ligand: Structure of designed TPR module-ligand complex. FEBS Journal *277*, 1058–1066.

Cyr, J.L., Pfister, K.K., Bloom, G.S., Slaughter, C.A., and Brady, S.T. (1991). Molecular genetics of kinesin light chains: generation of isoforms by alternative splicing. Proc. Natl. Acad. Sci. USA *88*, 10114–10118.

D'Andrea, L.D. and Regan L. (2003). TPR proteins: the versatile helix. Trends in Biochemical Sciences *28*, 655–662.

Das, A.K., Cohen, P.T., and Barford, D. (1998). The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein–protein interactions. The EMBO Journal *17*, 1192–1199.

DeLano W.L. (2010). The PyMOL Molecular Graphics System*,* version 1.5.1*,* Schrödinger, LLC*,* New York.

Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.-K., Shi, Y., and Yan, N. (2012). Structural basis for sequence-specific recognition of DNA by TAL effectors. Science *335*, 720–723.

Emsley, P., and Cowtan, K. (2004). *Coot*: model-building tools for molecular graphics. Acta Crystallographica Section D Biological Crystallography *60*, 2126–2132.

Edelhoch H. (1967) Spectroscopic determination of tryptophan and tyrosine in proteins. Biochemistry 6, 1948–1954.

Espagne, E., Lespinet, O., Malagnac, F., Da, C., Aury, M., Ségurens, B., Poulain, J., Anthouard, V., Grossetete, S., Khalili, H., et al. (2008). The genome sequence of the model ascomycete fungus Podospora anserina. Genome Biology *9*, R77.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Research *42*, D222–D230.

Fisher, S.Q., Weck, M., Landers, J.E., Emrich, J., Middleton, S.A., Cox, J., Gentile, L., and Parish, C.A. (2012). Evidence that the kinesin light chain domain contains tetratricopeptide repeat units. Journal of Structural Biology *177*, 602–612.

Frith, M.C., Saunders, N.F.W., Kobe, B., and Bailey, T.L. (2008). Discovering Sequence Motifs with Arbitrary Insertions and Deletions. PLoS Computational Biology *4*, e1000071.

Gindhart J.G. Jr, and Goldstein L.S.B. (1996). Tetratrico peptide repeats are present in the kinesin light chain. TIBS Letters 21, 52-53.

Headd, J.J., Echols, N., Afonine, P.V., Grosse-Kunstleve, R.W., Chen, V.B., Moriarty, N.W., Richardson, D.C., Richardson, J.S., and Adams, P.D. (2012). Use of knowledge-based restraints in *phenix.refine* to improve macromolecular refinement at low resolution. Acta Crystallographica Section D Biological Crystallography *68*, 381–390.

Hunter J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9(3), 90-95.

Johnson M.L., Straume M. (1994). Comments on the analysis of

sedimentation equilibrium experiments In T.M. Shuster, T.M. Laue, ed. (Modern Analytical Ultracentrifugation Boston: Birkhauser), pp. 37-65.

Kajander, T., Cortajarena, A.L., Main, E.R.G., Mochrie, S.G.J., and Regan, L. (2005). A New Folding Paradigm for Repeat Proteins. Journal of the American Chemical Society *127*, 10188–10190.

Kajava, A.V. (2002). What curves α-solenoids? Evidence for an α-helical toroid structure of Rpn1 and Rpn2 proteins of the 26 S proteasome. Journal of Biological Chemistry *277*, 49791–49798.

Karplus, P.A. and Diederichs, K. (2012). Linking crystallographic model and data quality. Science 6084, 1030-1033.

Kleywegt, G.J. (1996). Use of non-crystallographic symmetry in protein structure refinement. Acta Crystallographica Section D Biological Crystallography *52*, 842-857.

Kloss, E., Courtemanche, N., and Barrick, D. (2008). Repeat-protein folding: New insights into origins of cooperativity, stability, and topology. Archives of Biochemistry and Biophysics *469*, 83–99.

Kobe, B., and Kajava, A.V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends in Biochemical Sciences *25*, 509–515.

Lemaire, P.A., Lary, J., and Cole, J.L. (2005). Mechanism of PKR Activation: Dimerization and Kinase Activation in the Absence of Double-stranded RNA. Journal of Molecular Biology *345*, 81–90.

Main, E., Lowe, A., Mochrie, S., Jackson, S., and Regan, L. (2005). A recurring theme in protein engineering: the design, stability and folding of repeat proteins. Current Opinion in Structural Biology *15*, 464–471.

Main, E.R.G., Xiong, Y., Cocco, M.J., D'Andrea, L., and Regan, L. (2003). Design of Stable α-Helical Arrays from an Idealized TPR Motif. Structure *11*, 497–508.

Mak N.-S.M., Bradley P., Cernadas R.A., Bogdanove A.J., Stoddard B.L. (2012). The crystal structure of TAL Effector PthXo1 Bound to Its DNA Target. Science 335, 716-719.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). *Phaser* crystallographic software. Journal of Applied Crystallography *40*, 658–674.

Mello, C.C., and Barrick, D. (2004). An experimentally determined protein folding energy landscape. Proc. Natl. Acad. Sci. USA *101*, 14102–14107.

Mosavi, L.K., Minor, D.L., and Peng, Z. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. Proc. Natl. Acad. Sci. USA *99*, 16029–16034.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallographica Section D: Biological Crystallography *53*, 240–255.

van Ooijen, G., Mayr, G., Kasiem, M.M.A., Albrecht, M., Cornelissen, B.J.C., and Takken, F.L.W. (2008). Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. Journal of Experimental Botany *59*, 1383–1397.

Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods in Enzymol. 276, 307–326.

Pace, C.N. (1986). Determination and analysis of urea and guanidine hydrochloride denaturation curves. Methods in Enzymology *131*, 266–280.

Pape, T., and Schneider, T.R. (2004). *HKL2MAP*: a graphical user interface for macromolecular phasing with *SHELX* programs. Journal of Applied Crystallography *37*, 843–844.

Parmeggiani, F., Pellarin, R., Larsen, A.P., Varadamsetty, G., Stumpp, M.T., Zerbe, O., Caflisch, A., and Plückthun, A. (2008). Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core. Journal of Molecular Biology *376*, 1282–1304.

Pernigo S., Lamprecht A., Steiner R.A., Dodding M.P. (2013). Structural basis for Kinesin-1: cargo recognition. Science 340, 356-359

Sanner, M.F., and Olson, A.J. and Spehner J.-C. (1996). *Reduced Surface: an Efficient Way to Compute Molecular Surfaces.* Biopolymers 38, 305-320.

Schuck, P. (2000). Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. Biophysical Journal *78*, 1606–1619.

Schneider, T.R. and Sheldrick, G.M. (2002). Substructure Solution with SHELXD. Acta Crystallographica Section D Biological Crystallography 58, 1772-1779.

Sheldrick, G.M. (2010). Experimental phasing with *SHELXC* / *D* / *E*: combining chain tracing with density modification. Acta Crystallographica Section D Biological Crystallography *66*, 479–485.

Sikorski R.S., Boguski M.S., Goebl M., and Heiter P. (1990). A repeating amino acid motif in CDC23 defines a new family of proteins and a new relationship among genes required for mitosis and RNA synthesis. Cell 60(2), 307-317.

Stafford, W.F., and Sherwood, P.J. (2004). Analysis of heterologous interacting systems by sedimentation velocity: curve fitting algorithms for estimation of sedimentation coefficients, equilibrium and kinetic constants. Biophysical Chemistry *108*, 231–243.

Street, T.O., Courtemanche, N., and Barrick, D. (2008). Protein Folding and Stability Using Denaturants. In Methods in Cell Biology, (Elsevier), pp. 295–325.

Terwilliger, T.C. (2003). SOLVE and RESOLVE: automated structure solution and density modification. Methods Enzymol. 374, 22–37.

Urvoas, A., Guellouz, A., Valerio-Lepiniec, M., Graille, M., Durand, D., Desravines, D.C., van Tilbeurgh, H., Desmadril, M., and Minard, P. (2010). Design, Production and Molecular Structure of a New Family of Artificial Alpha-helicoidal Repeat Proteins (αRep) Based on Thermostable HEAT-like Repeats. Journal of Molecular Biology *404*, 307–327.

Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K., and Plückthun, A. (2008). Folding and Unfolding Mechanism of Highly Stable Full-Consensus

Ankyrin Repeat Proteins. Journal of Molecular Biology *376*, 241–257.

Wheeler, T.J., Clements, J., and Finn, R.D. (2014). Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics *15*, 7.

Wowor, A.J., Yu, D., Kendall, D.A., and Cole, J.L. (2011). Energetics of SecA Dimerization. Journal of Molecular Biology *408*, 87–98.

Xu, Y. (2004). Characterization of macromolecular heterogeneity by equilibrium sedimentation techniques. Biophysical Chemistry *108*, 141–163.

Yan, N., Chai, J., Lee, E.S., Gu, L., Liu, Q., He, J., Wu, J.-W., Kokel, D., Li, H., Hao, Q., et al. (2005). Structure of the CED-4–CED-9 complex provides insights into programmed cell death in Caenorhabditis elegans. Nature *437*, 831–837.

Zhu, H., Lee, H.Y., Tong, Y., Hong, B.-S., Kim, K.-P., Shen, Y., Lim, K.J., Mackenzie, F., Tempel, W., and Park, H.-W. (2012). Crystal Structures of the Tetratricopeptide Repeat Domains of Kinesin Light Chains: Insight into Cargo Recognition Mechanisms. PLoS ONE *7*, e33943.

CHAPTER 4

A nearest neighbor analysis of a naturally occurring repeat protein with high internal sequence identity

4.1 Abstract

Repeat proteins are formed from linear arrays of modular structural units. Due to their low contact order compared to globular proteins, they are able to tolerate the addition and removal of whole repeats and still remain folded in solution. This feature has enabled their equilibrium unfolding to be described using nearest-neighbor (Ising) models. Globally analyzing a series of equilibrium unfolding data using these models enables a thermodynamic description of cooperativity. This approach is generally restricted to designed consensus versions of repeats, as high sequence identity between adjacent repeats (internal sequence identity), is often required for modeling systems. Here we present a heteropolymer analysis of a natural 42PR repeat protein system from the fungus *P. anserina* (*Pa* 42PRs) using whole and half repeats. For whole repeat analysis, we find 42PRs to have increased magnitudes of both ΔG_i and $\Delta G_{i,i+1}$ terms compared to c34PRs, consistent with the high level of apparent cooperativity observed in previous studies. Analyzing *Pa* 42PRs using the single helix approach from Chapter 2, reveals the stability of the B_i:A_{i+1} interface is insufficient to completely counteract the instability of the A-helix. To understand if there were any key structural features within this interface, we examined the PDB and discovered a conserved hydrogen bond with many nPR B_i:A_{i+1} interfaces. Elimination of the donor group (Y16F) results in significant

unfolding of the terminal repeats of *Pa* 42PR arrays, suggesting this H-bond confers significant interfacial coupling energy between repeats. Taken together, these results provide important details regarding the mechanism of stabilization through consensus design, and point to key metrics to be used for design of a 42PRs consensus sequence.

4.2 Introduction

Cooperativity in protein folding is characterized by a lack of intermediate populations at equilibrium (Barrick, 2009; Chan et al., 1995). Repeat proteins are composed of sequentially ordered domains which stack to form elongated structures (Kajava, 2001; Kloss et al., 2008; Main et al., 2005). In contrast to globular proteins, the low contact order of repeat proteins enables them to tolerate addition and removal of structural units without compromising the fold of the molecule. Recently, the folding mechanisms of consensus designed repeat proteins have been studied using one dimensional Ising models (Aksel et al., 2011; Kajander et al., 2005; Mello and Barrick, 2004; Wetzel et al., 2008). These studies show that nearest neighbor interactions couple adjacently folded repeats, resulting in cooperative folding. This analysis yields two (or more, Chapter 2) energetic terms ΔG_i, and $\Delta G_{i,i+1}$, which describe the free energy of folding one repeat, and interactions between repeats, respectively.

Although these models provide significantly more insight than two-state analysis, Ising approaches are typically limited to designed consensus proteins. The requirement of identical repeats has prevented the study of many natural proteins using these models, due to their low (~25%) seqeunce identity between repeats. Therefore, there has not been a detailed energetic comparison between natural and designed repeat proteins using these models.

Designed α-helical repeat proteins tend to be considerably more stable than their natural counterparts (Aksel et al., 2011; Kajander et al., 2005; Kloss et al., 2008; Urvoas et al., 2010). Structurally, it is difficult to pinpoint molecular interactions that could be contributing favorably to stability. In the leucine rich repeat protein YopM, adjacent repeats contain oppositely charged residues, YopM's stability has a peculiarly high salt dependence (Kloss and Barrick, 2008) and, interestingly, YopM also increases in stability upon *removal* of specific internal repeats (Vieux and Barrick, 2011), and therefore there are multiple interconnected factors which contribute to global stability.

In addition, structural and sequence differences do not always result in energetic differences as we saw in Chapter 2 (A_i:B_i and B_i:A_{i+1} interfaces). It would be of considerable interest to be able to visualize favorable interactions in protein structures, as this has many potential applications in bioengineering (Cunha et al., 2013; Grove et al., 2010; Main et al., 2005) and general protein design. In addition, the simple observation of stabilization through consensus deisng is intriquing. What causes consensus desined proteins to be more stable than their natural counterparts? When viewed from a cooperativity standpoint, are intrinsic units stabilized more than interfacial interactions through consensus design? These questions and others have not been addressed due to our inability to determine precise energetics of natural systems for

comparison. While modeling the naturally occuring notch Ankyrin domain repeats using a single energy term was successful (Mello and Barrick, 2004), this does not allow the precise resolution of the energetics of individual repeats—just their average.

The *Pa* 42PRs presented in Chapter 3 provide an excellent system to delve into these questions. Therefore, we sought to apply a nearest-neighbor approach to *Pa* 42PRs, using whole repeats as well as single helices as was done with c34PRs in Chapter 2. We were able to model Pa42PRs using a set of nearest neighbor parameters for whole repeats and single helices. To begin to probe the structural determinants of our observed energetics, we targeted the hydrogen bond found in our structure from Chapter 2. Analysis of the worldwide PDB revealed this to be a conserved motif among nPR families. Elimination of this hydrogen bond donor in *Pa* 42PRs results in multi-repeat unfolding, which is consistent with its role of conferring significant stability in *Pa* 42PR interfaces. Taken together, our nearest neighbor and mutational analyses have implications for nPR consensus design by identifying general and specific mechanisms of stabilization.

4.3 Results

Whole-repeat homopolymeric Ising analysis of *Pa* 42PRs

To obtain a more mechanistic understanding of the apparent increase in cooperativity in *Pa* 42PRs compared to c34PRs observed in Chapter 3, we analyzed a set of *Pa* 42PR constructs using a one-dimensional Ising model (Aksel and Barrick, 2009; Aksel et al., 2011; Kajander et al., 2005; Mello and Barrick, 2004), and compared them to a similar analysis of c34PR constructs.

We globally fit all *Pa* 42PR unfolding transitions from Chapter 3 to Ising model (Figure 4.1, closed circles and solid lines) using one intrinsic and one interfacial free energy term, and compared them to a separate global fit of c34PR unfolding transitions (Figure 4.1, open circles and dashed lines). The global parameters obtained from these fits are shown in Table 4.1. Cooperativity in *Pa* 42PRs arises from unfavorable intrinsic (ΔG_i) repeat folding, and from favorable interfacial ($\Delta G_{i,i+1}$) coupling between adjacent folded repeats. The cooperativity enhancement observed for *Pa* 42PRs results from an increase in magnitude of both ΔG_i and $\Delta G_{i,i+1}$ compared to c34PRs.

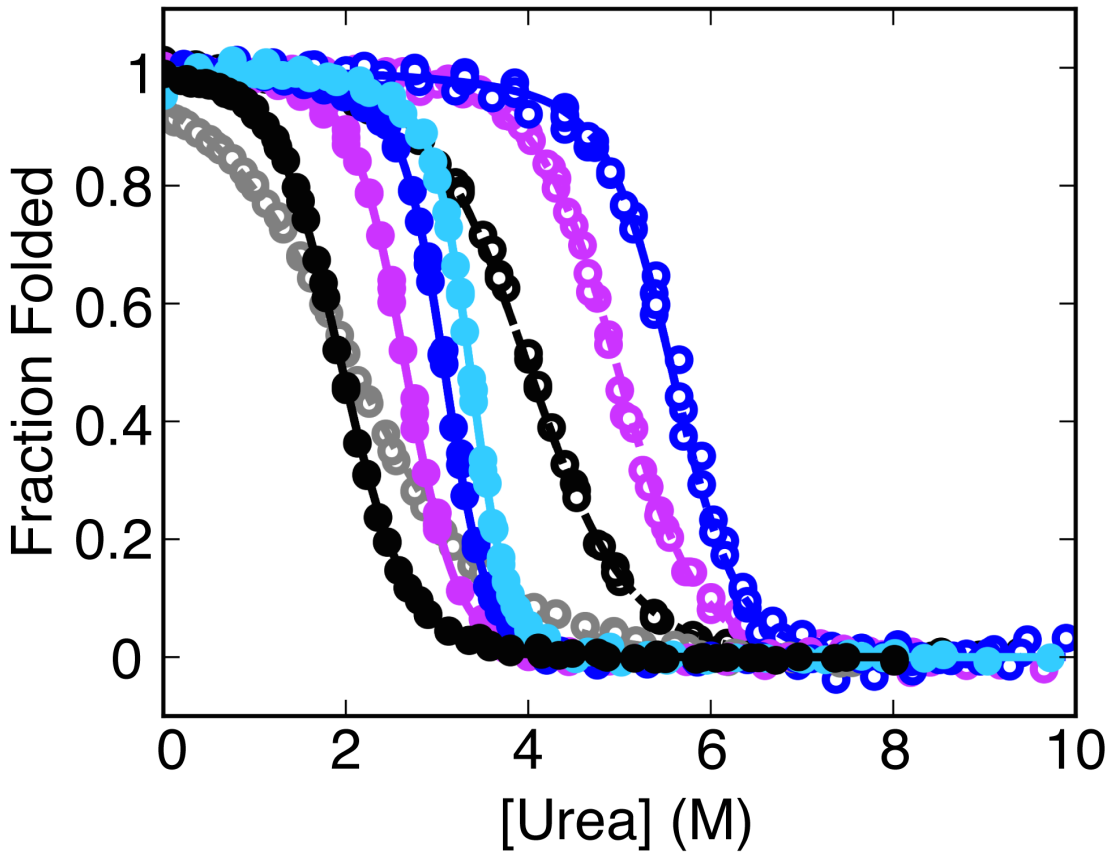


Figure 4.1 Global fits of normalized equilibrium unfolding transitions of *Pa* 42PRs and c34PRs to one dimensional Ising models. Closed circles show *Pa* 42PR constructs, and are colored as in Figure 3.4, ranging from three to five total repeats. Open circles show c34PR constructs: B(AB)S (grey), $B(AB)_2S$ (black), $B(AB)_3S$ (purple), and $B(AB)_4S$ (blue). Solid and dashed lines result from globally fitting one-dimensional Ising models to *Pa* 42PRs and c34PRs, respectively. Intrinsic energies were modeled using a single parameter for *Pa* 42PRs ($N_AB/AB/AC_B$ repeats) and c34PRs (BA/BS repeats).

Table 4.1. Homopolymer thermodynamic Ising parameters for *Pa* 42PR and c34PRs

Repeat	χ^2/ν[a]	ΔG_i[b]	$\Delta G_{i,i+1}$[b]	m_i[c]
Pa 42PRs	$6.5E^{-5}$	2.01± 0.03 (1.81,2.2)[d]	-4.63 ± 0.04 (-4.97,-4.38)[d]	-0.57 ± 0.004 (-0.6,-0.54)[d]
c34PRs	$1.64E^{-4}$	1.39 ± 0.04 (1.05,1.73)[d]	-4.3 ± 0.07 (-4.93,-3.83)[d]	-0.38 ± 0.005 (-0.43,-0.35)[d]

Ising parameters were obtained from a global fit of a nearest-neighbor model to *Pa* 42PR and c34PR equilibrium unfolding curves for constructs with integral numbers of repeats (Figure 4.1). Three or more independent unfolding transitions for each construct were included.

[a] Reduced chi-squared

[b] $kcal*mol^{-1}$

[c] $kcal*mol^{-1}*M^{-1}$

[d] 95% confidence intervals shown in parenthesis were obtained by a rigorous F-statistics method (Johnson and Straume, 1994), and are compared to the estimations from the covariance matrix corresponding to each fit.

Design of *Pa* 42PRs lacking N- or C-terminal capping repeats

To determine whether the energies of N_A, AB, and AC_B repeats were identical, we created constructs that have terminal cap deletions (Figure 4.2). We refer to these constructs using the NRC notation used in cANK repeat protein folding (Aksel et al., 2011).

These constructs were created in a manner as similarly described in Chapter 3. The NRC constructs studied here display α-helical spectra with similar shapes to constructs studied in Chapter 3. These NRC constructs are moderately well-behaved in solution at low concentrations as assessed by sedimentation velocity analytical ultracentrifugation studies (data not shown), and display similar hydrodynamic properties as the doubly capped constructs studied in Chapter 3.

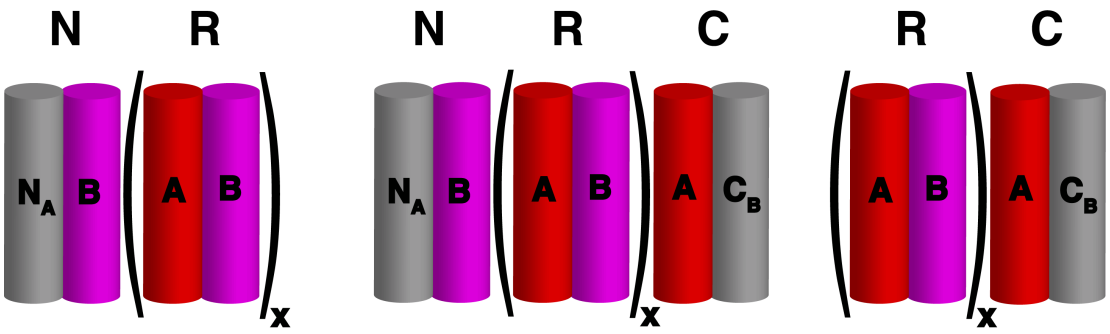


Figure 4.2. *Pa* 42PR constructs for NRC heterpolymeric Ising analysis. N, R, and C repeats contain native (R) or substituted (N,C). *Pa* 42PR A and B-helices. N and C substitutions are outlined in Figure 3.2.

Urea-induced equilibrium unfolding of the NRC series

To measure the stability of our constructs, we performed urea-induced equilibrium unfolding studies. *Pa* 42PRs tolerate the removal of one capping repeat very well, retaining apparent two-state equilibrium unfolding (Figure 4.3, solid circles), showing the same increase in steepness with repeat number as described in Chapter 3. While most constructs show a similar level of destabilization upon removal of N, R, or C repeats, there is a clear distinction between NR3, and R3C, with R3C being reproducibly more stable. This suggests the polar substitutions made on the C-terminal repeat are stabilizing (Figure 3.2).

Whole-repeat heteropolymeric Ising analysis of *Pa* 42PRs

To determine the energetic parameters of the N, R, and C repeats, we globally fit our *Pa* 42PR unfolding titration data using a model similar to that used for cANKs. By studying a length dependence of different constructs with different capping structures, we can uniquely determine the energy terms ΔG_N, ΔG_R, ΔG_C, and $\Delta G_{i,i+1}$, which describe intrinsic and interfacial interactions, respectively, in *Pa* 42PRs.

We created six (M1-M6) nearest-neighbor models and fit to our data with a combinatorial set of energetic parameters. As was the case with global fitting in Chapter 2, our data were fit well by all models (Figure

4.3). A summary of the fit qualities from each model and their associated best-fit energy parameters is shown in Table 4.2

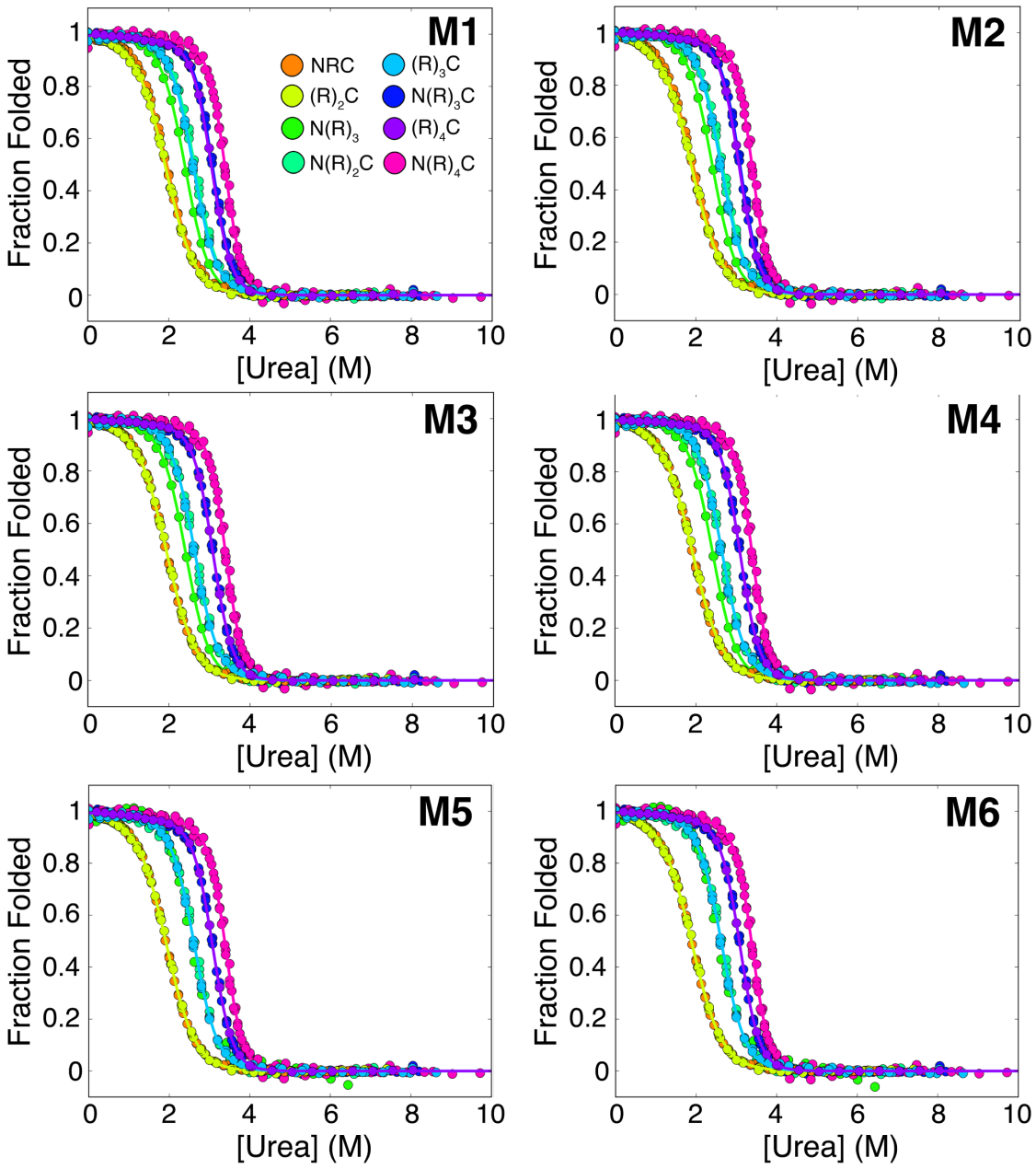


Figure 4.3. Global fit of Models M1-M6 to the *Pa* 42PR NRC series. Best-fit parameters and fit statistics for this model can be found in Table 4.1 Construct colors are displayed in the M1 figure legend and are the same for M1-M6.

Table 4.2. *Pa* 42PR NRC whole repeat heteropolymer nearest-neighbor model analysis

Model	$\chi 2/\nu$	ΔG_N [a]	ΔG_R [a]	ΔG_C [a]	$\Delta G_{i,i+1}$ [a]	m_i [b]	$m_{i,i+1}$ [a]
M1	9.74E-5	2.05 (0.87,3.21)	1.97 (0.95,3.28)	1.44 (0.46,2.7)	-4.32 (-6.05,2.98)	-0.78 (-1.13,-0.73)	0.31 (-0.27,0.25)
M2	1.01E-4	2.58 (2.22,2.944)	2.65 (2.29,3.01)	2.11 (1.9,2.33)	-5.23 (-5.67,-4.8)	-0.54 (-0.58,-0.51)	0
M3	1.01E-4	ΔG_R	2.08 (2.08,3.29)	1.44 (0.44,2.74)	-4.31 (-6.1,-2.93)	-0.77 (-1.13,-0.73)	0.3 (-0.29,-0.25)
M4	1.05E-4	ΔG_R	2.61 (2.26,2.97)	2.09 (1.88,2.32)	-5.16 (-5.65,-4.78)	-0.54 (-0.58,-0.51)	0
M5	1.44E-4	ΔG_R	1.51	ΔG_R	-3.92	-0.78	0.304
M6	1.49E-4	ΔG_R	2.16	ΔG_R	-4.78	-0.55	0

[a] $kcal*mol^{-1}$

[b] $kcal*mol^{-1}*M^{-1}$

Parenthesis indicate 95% confidence intervals obtained by F-statistics

To determine the correct model to interpret parameters, we performed F-statistic comparison tests, as was done in Chapter 2 (Table 4.3). We find models M1-M4 to fit the data well. While M1 has the lowest reduced χ^2, it has the greatest number of parameters, and the F-statistic comparisons reveal our confidence in these parameters is very close to the 1σ limit, compared to M2-M4. Therefore, we cannot conclude with high confidence it is the more appropriate one to use. Instead, we prefer model M2, as it contains fewer parameters (Figure 4.2). Model M3 is also reasonable, as it eliminates the necessity to fit a unique N repeat energy, due to its similarity to ΔG_R. Best-fit parameters and their associated errors are displayed in Table 4.2.

Table 4.3. *Pa* 42PR NRC heteropolymer nearest neighbor model F-statistic comparison

Model	M1	M2	M3	M4	M5	M6
M1						
M2	1.04 (66)					
M3	1.04 (66)	1.0 (50)				
M4	1.08 (80)	1.03 (70)	1.035 (67)			
M5	1.47[a]	1.42[a]	1.41[a]	1.37[a]		
M6	1.53[a]	1.47[a]	1.48[a]	1.42[a]	1.035 (67)	

[a] An F > 1.34 corresponds to a probability > 99.9 %.

Due to similarity in degrees of freedom between the models, the critical F values for 1σ, 2σ, and 3σ correspond to ~1.035, ~1.17, and ~1.25, respectively, for all model comparisons. Values in parenthesis indicate the % confidence in the better-fit model from each comparison. F-ratios are calculated as column/row.

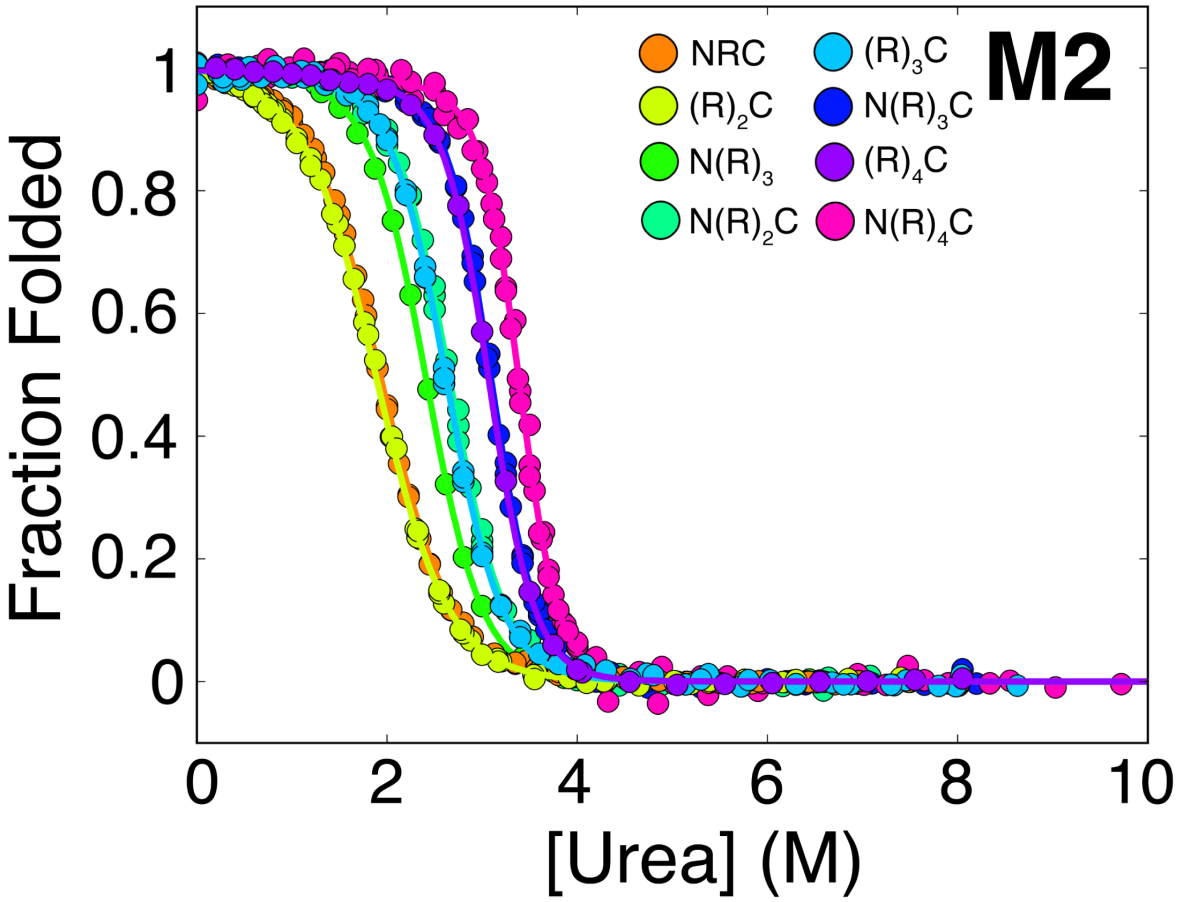


Figure 4.4. Global fit of model M2 *Pa* 42PR to the NRC series. Best-fit parameters and fit statistics for this model can be found in Table 4.3.

Single helix heteropolymeric Ising analysis of *Pa* 42PRs

In Chapter 2, I presented a framework to analyze both intra- and inter-repeat coupling energies, in addition to single α-helical energies in cTPRs. As *Pa* 42PRs belong to the same superfamily and have a similar structure (Table 3.4 and Figure 3.6), I wanted to see if the same approach could be taken with *Pa* 42PRs. To do this, I created constructs which altered single helices of *Pa* 42PRs, as outlined in Figure 2.4 (c34PRs).

We were able to create, express, purify, and study these constructs in urea-induced equilibrium unfolding and CD spectroscopy studies. CD spectra for representative NRC series constructs, and those altering single helices are displayed in Figure 4.4. These constructs all have similar spectral shapes, but their magnitudes are different. This is likely to be due to our limited ability to determine accurate protein concentrations due to their low extinction coefficients. This hypothesis is consistent with the unrealistic MRE values displayed on the y-axis.

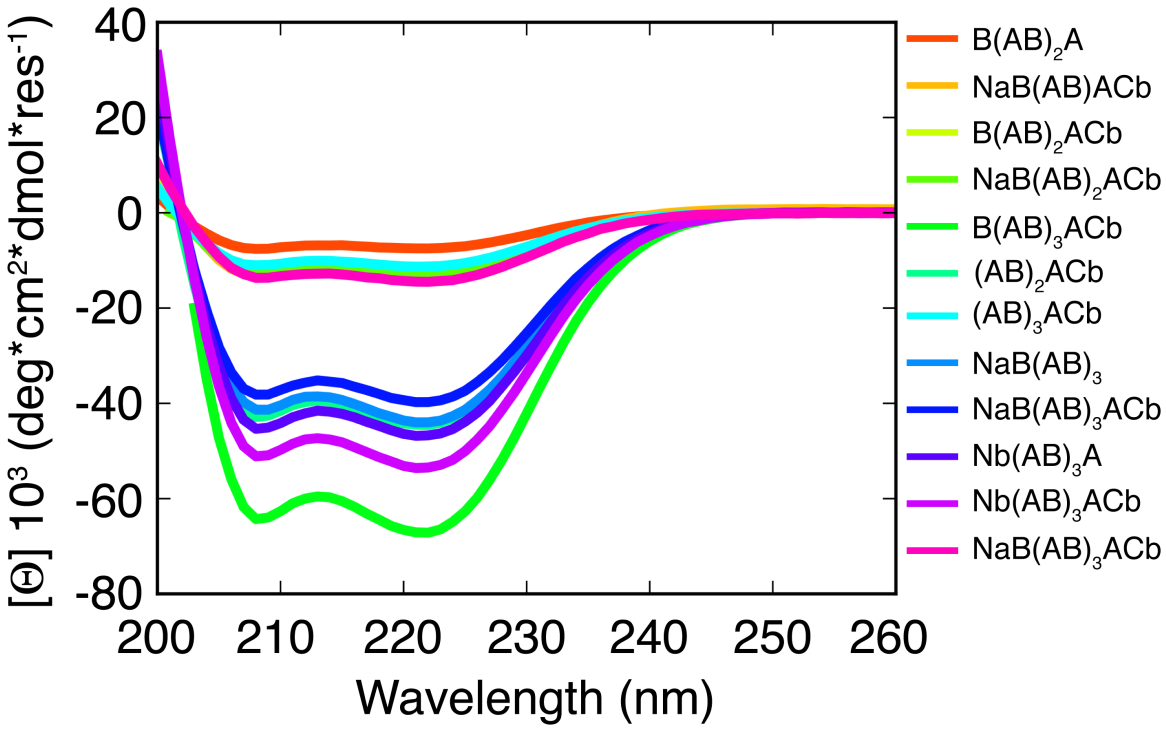


Figure 4.5. Representative Far-UV CD spectra of *Pa* 42PR constructs.

Ising analysis treating individual helices of *Pa* 42PRs

Although we were able to obtain these *Pa* 42PRs (by varying single helices) in solution, measure equilibrium unfolding transitions, and fit them well using single helix heteropolymeric Ising models, the fitted parameters show very large correlation. This correlation appears to result from the low stability of the A-helix on the C-terminus. An example of a fit containing these constructs is shown in Figure 4.6. The parameter correlation can be visualized by the fitted curves, where constructs containing C-terminal A-helices are baseline adjusted to a value corresponding to the fractional contribution of that helix in the array (for example a construct with ten total helices and an A-helix on the C-terminus, would have a baseline adjusted y-intercept at 0.9). In addition, the correlation can be directly observed in parameter space by analyzing parameters obtained through bootstrapping iterations.

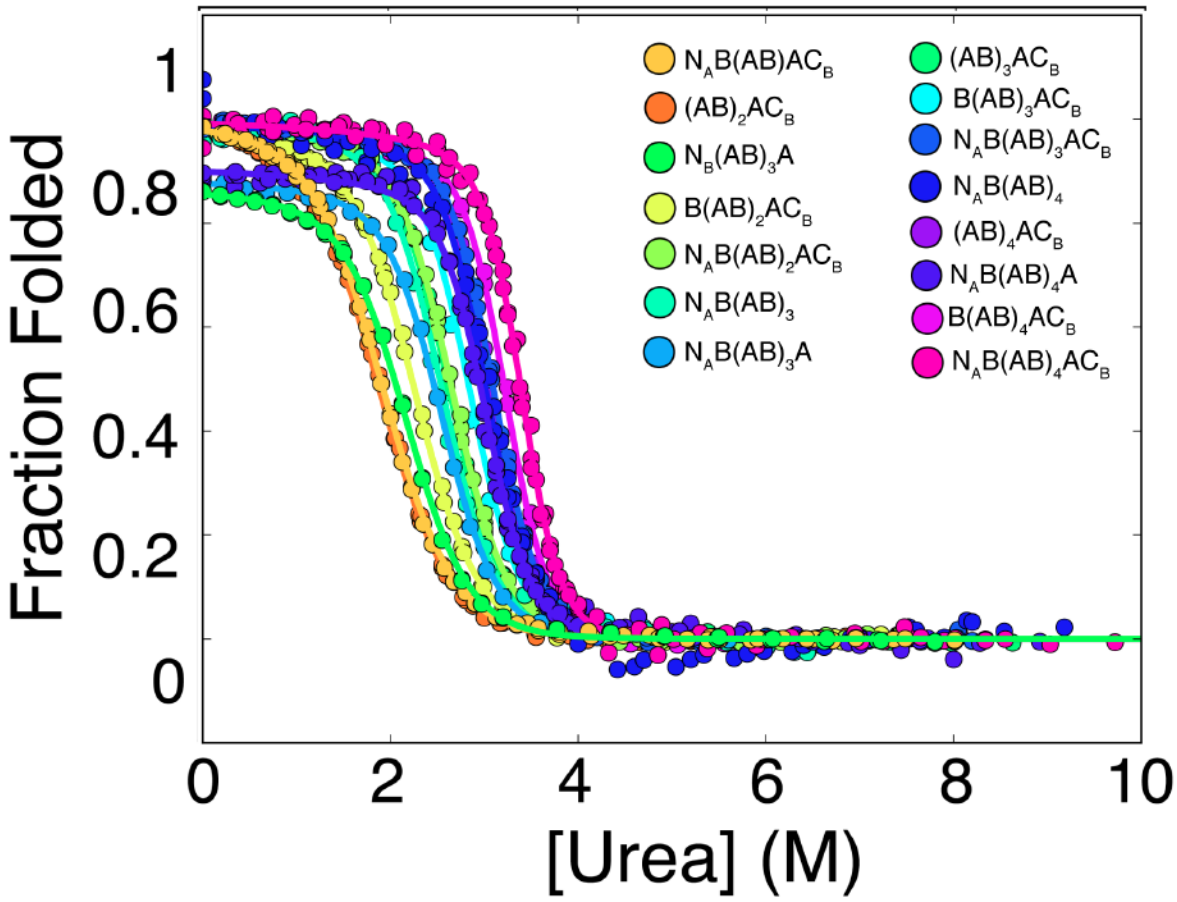
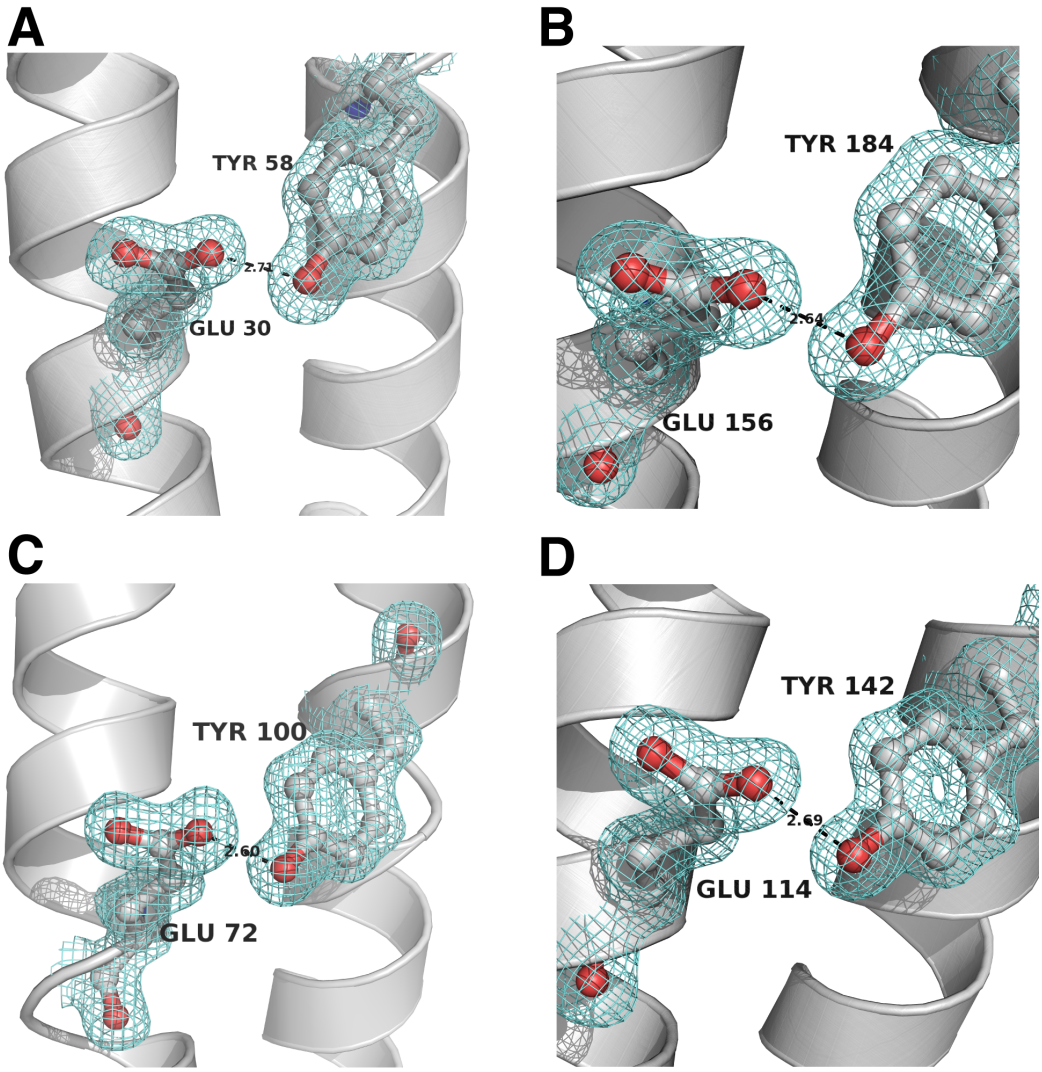


Figure 4.6. Global fit of *Pa* 42PRs with single helix additions. This fit has parameter correlation and extreme best-fit energy values as a result.

A conserved hydrogen bond in the nPR family

In the structure of the *Pa* 42PR we determined in Chapter 3, there is a regular Tyr OηH--$^{-}$OεC Glu hydrogen bond in the interface between all five repeats. The electron density for the 4Y6W hydrogen bond is very well defined at all four (B_i:$A_{i,i+1}$) interfaces (Figure 4.7). The Tyr residues are at position 16 in the 42PR motif on the A-helix. In the interaction, it flips back across the interface to engage in an H-bond interaction with the Glu located within the B-helix of the previous repeat. Due to the sequence and structure conservation, we hypothesized the Tyr may contribute substantial stability to the interface. Most of the interfacial interactions in nPRs are hydrophobic (Figure 3.6), and the low polarity of this environment could increase the strength of a hydrogen bond interaction (Gao et al., 2009). Interestingly, we have observed this interaction motif in numerous TPR/nPR structures (Figure 4.8).



.

Figure 4.7. (A-D) Tyr OηH--ΟεC Glu hydrogen bonds in 4Y6W**.** In all interactions, the hydrogen bonding distance is 2.6-2.7 Å.

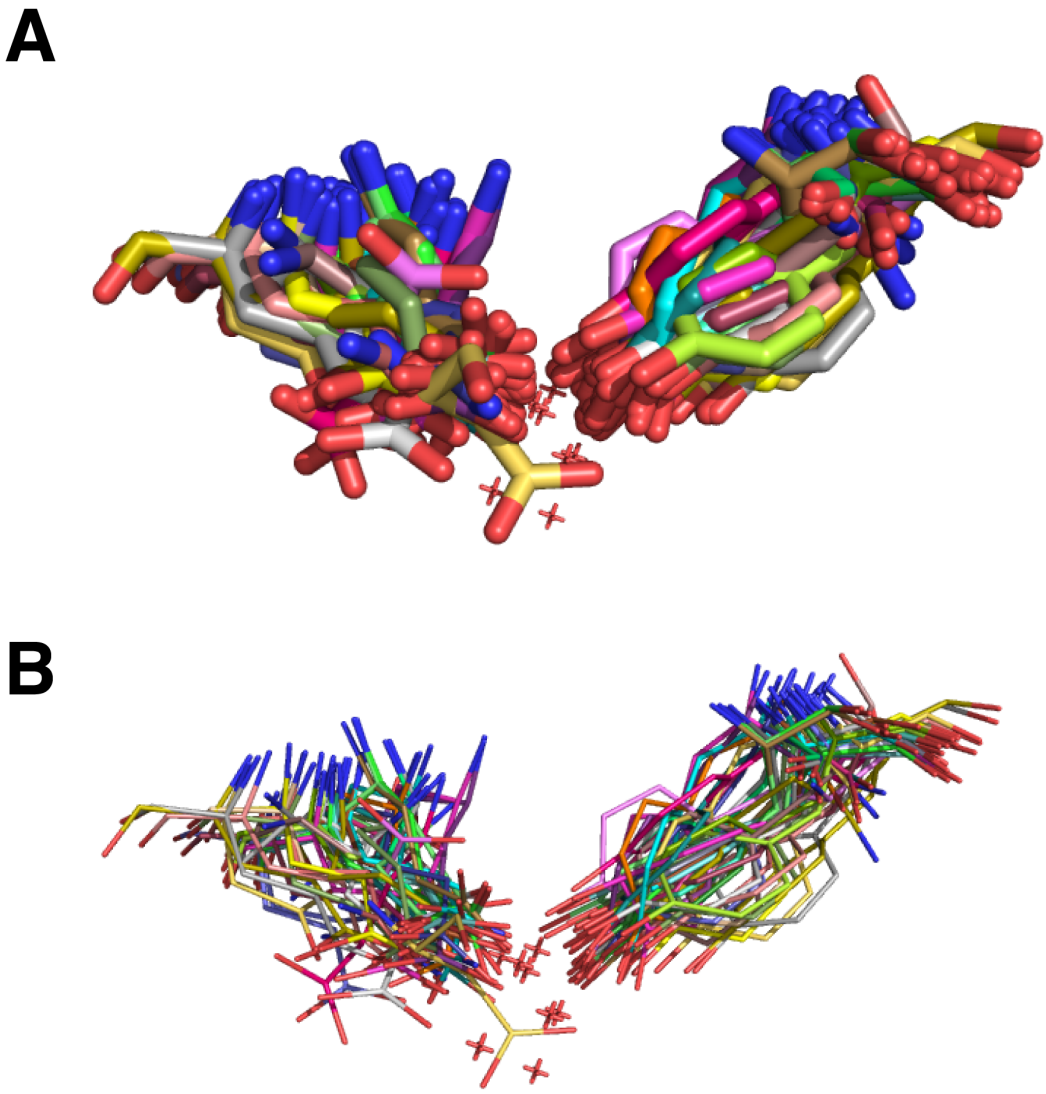


Figure 4.8. (A-B) Tyr OηH--ΟεC Glu or Nη Gln hydrogen bonds high-resolution PDB structures of TPR motifs**.**

To probe this interaction, we made Y16F substitutions to single A-helices. By comparing the effect of inter-repeat Y16F substitutions, to effects of substitution of Y16F on the N-terminus, we could assess the intrinsic effect of single Y16F substitutions alone, and when perturbing a hydrogen bond (Figure S4.2). We used the construct $B(AB)_4AC_B$ (Figure 4.6) as a reference point for engineering substitutions. We introduced a Y16F substitution at the interface between the third and fourth helices from the N-terminus to create $BAB^*(AB)_3AC_B$, where * indicates a Y16F substitution. We performed urea-induced equilibrium unfolding experiments on these constructs and fit them using a two-state unfolding model.

We found the introduction of a Y16F substitution in this context to be destabilizing by ~1.3 kcals/mol, with a slight (0.1 kcals/mol*M) decrease in m-value compared to $B(AB)_4AC_B$ (Figure 4.8 and Table 4.4). If the N-terminal BAB unit were to completely unfold, we would expect a free energy similar to that of $(AB)_3AC_B$, which is ~3 kcals/mol less stable than $B(AB)_4AC_B$. Therefore, while the hydrogen bond does not confer complete interfacial stability (BAB is not able to fold in isolation), it contributes significant stability to the interface. Moreover, the lack of multistate unfolding for $BAB^*(AB)_3AC_B$, coupled with the fact its m-value is minimally perturbed, suggests the size of the cooperative unit is unaffected by the

substitution despite this significant destabilization. This also suggests the effect of the substitution is felt along the entire molecule.

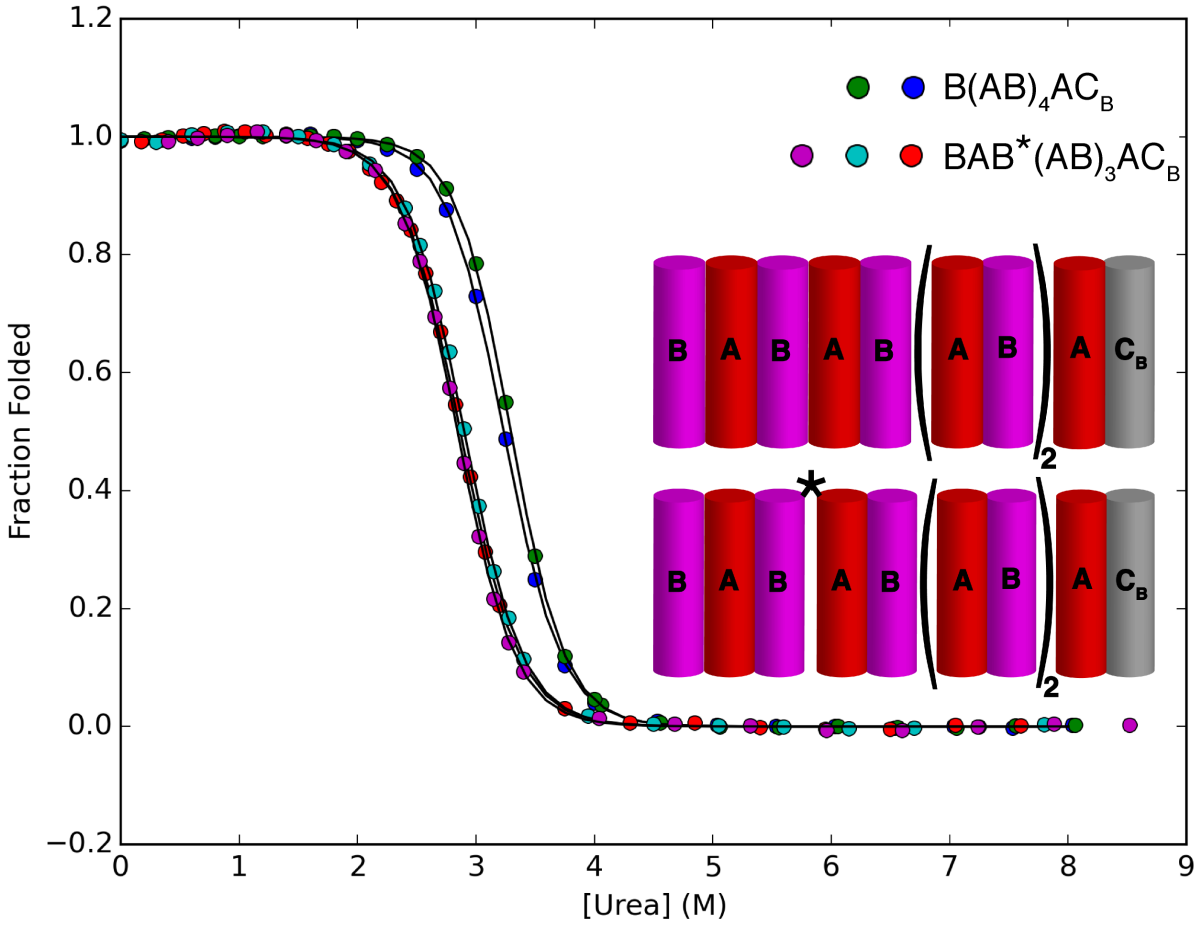


Figure 4.9. (A-B) Tyr OηH--ˉOεC Glu hydrogen bonds perturbation. The dots represent multiple experiments for each construct. The asterisk represents the position of the hydrogen bond disruption, and cartoon representations describing helices are shown to enhance interpretation.

Table 4.4. WT and Y16F *Pa* 42PR construct two-state parameters

Construct	ΔG°_{H20}[a]	M-value[b]
$(AB)_3AC_B$	5.22 ± 0.17	2.02 ± 0.04
$(*AB)(AB)_2AC_B$	5.5	2.1
$B(AB)_4AC_B$	8.19 ± 0.3	2.51 ± 0.07
$BAB*(AB)_3AC_B$	6.89 ± 0.09	2.4 ± 0.04

Parameters were obtained from two-state fits of *Pa* 42PR constructs. Errors represent standard deviations of the mean. * indicates a the position of a Y16F substitution.
[a] $kcal*mol^{-1}$
[b] $kcal*mol^{-1}*M^{-1}$

4.4 Discussion

In this study we have successfully applied nearest-neighbor models to a natural repeat protein to uniquely determine the thermodynamic aspects of cooperativity in a naturally occurring repeat protein. We find that, compared to c34PRs, *Pa* 42PRs are less stable, based on midpoints, but are more cooperative. This cooperativity enhancement is characterized by an increased magnitude of both intrinsic repeat folding (more unfavorable terms) and interfacial coupling (increased stability of interfacial interactions.

Collectively, this leads to a different picture of the folding landscape, as intermediates are less likely to form in *Pa* 42PR unfolding transitions. In contrast, the 42PR array is still less cooperative than consensus ankyrins, which have further increased intrinsic free energy magnitudes (~4 kcals/mol greater) than *Pa* 42PRs. Therefore, *Pa* 42PRs represent intermediate positions along the cooperativity scale observed in repeat protein systems (c34PRs < *Pa* 42PRs < cANKs < cLRRs).

It is unfortunate we were not able to resolve the intrinsic stabilities of each helix in *Pa* 42PRs as we were for c34PRs in Chaper 2. The parameter correlation observed between ΔG_A and $\Delta G_{Ai:Bi}$ is likely a result of insufficient $\Delta G_{Bi:Ai+1}$ interface energy to overcome the instability of the A-helix in *Pa* 42PRs. Because the A-helix energy is not determined, it's value floats to an arbitrary number, and the rest of the energetic

parameters must shift accordingly to compensate. Despite this, we can estimate the stability of the B helix by combining two-state global energetics and parameters obtained from Ising analysis of whole-repeats. Using this approach, we estimate the stability of the B-helix to be ~4 kcals/mol. Therefore, by estimation, the c34PR B-helix is nearly 3 kcals/mol more stable than the *Pa* 42PR B-helix.

The effect of the conservative substitution Y16F to eliminate a hydrogen bond in the interfacial region of *Pa* 42PRs is quite striking. Minimal perturbation by removal of a hydroxyl group results in a stability decrease of ~1.3 kcal/mol. The interfacial interaction energy within an interface is ~5 kcal/mol which is offset by the intrinsic folding of neighboring repeats on the order of ~2.3 kcal/mol (Table 4.2). In the context of the substitution $B(AB)_4AC_B$ and $BAB^*(AB)_3AC_B$, Figure 4.9), the hydrogen bond perturbation does not result in complete unfolding of the N-terminal BAB unit. Still, it is apparent that this conserved hydrogen bond in nPRs (Figure 2.7) is likely to be a contributing factor in determining inter-repeat coupling.

Since both 42PRs and c34PRs belong to the same family (nPRs), a comparison of their Ising parameters can provide estimates into the mechanism of stabilization through consensus design. Because interfacial energies are stronger in *Pa* 42PRs, it suggests that consensus design stabilizes intrinsic repeat units. This makes sense, as secondary structural

elements are more highly conserved in repeat proteins, and specific hydrogen bonds and interactions across interfaces are less likely to be widespread. However, to determine if these stabilization trends are a result of consensus design, or a result of motif length, a similar analysis would need to be performed on the consensus 42PR sequence (Figure 2.7).

It has been a major challenge in protein folding and biophysics to understand how energy in systems is subdivided. The nPR systems studied in this work have proven to be excellent systems to understand the intricate relationships between structure, sequence, and stability in protein folding. The extended Ising model approach in Chapter 2 provided a route to measure energetics of individual helices, and I am excited about the prospect of this methodology to be used in other contexts outside of protein folding.

4.5 Experimental Procedures

All common experimental procedures were carried out as previously described in Chapter 3.

4.6 Supplemental information

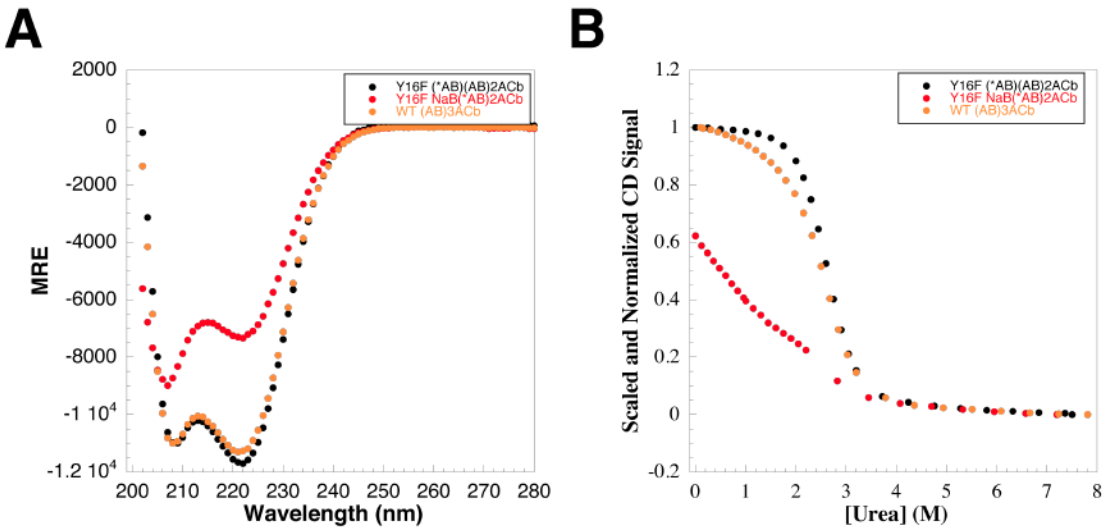


Figure S4.1. Effects of hydrogen bond disruption in two contexts. (A) CD spectra of WT $(AB)_3AC_B$, (Gold) Y16F $*(AB)_3AC_B$ (Black), and Y16F $N_AB*(AB)_2AC_B$ (Red), where the asterisk indicates the A-helix containing the Y16F substitution. In Y16F $*(AB)_3AC_B$, there is no h-bond disrupted, as the F is solvent exposed in this context.

4.7 References

Aksel, T., and Barrick, D. (2009). Chapter 4 Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models. In Methods in Enzymology, (Elsevier), pp. 95–125.

Aksel, T., Majumdar, A., and Barrick, D. (2011). The Contribution of Entropy, Enthalpy, and Hydrophobic Desolvation to Cooperativity in Repeat-Protein Folding. Structure *19*, 349–360.

Chan HS, Bromberg S, Dill KA. Models of cooperativity in protein folding Phil. Trans. R. Soc. Lond. B April 1995 Volume: 348 Issue: 1323

Cunha, E.S., Hatem, C.L., and Barrick, D. (2013). Insertion of Endocellulase Catalytic Domains into Thermostable Consensus Ankyrin Scaffolds: Effects on Stability and Cellulolytic Activity. Applied and Environmental Microbiology *79*, 6684–6696.

Grove, T.Z., Osuji, C.O., Forster, J.D., Dufresne, E.R., and Regan, L. (2010). Stimuli-Responsive Smart Gels Realized via Modular Protein Design. Journal of the American Chemical Society *132*, 14024–14026.

Kajander, T., Cortajarena, A.L., Main, E.R.G., Mochrie, S.G.J., and Regan, L. (2005). A New Folding Paradigm for Repeat Proteins. Journal of the American Chemical Society *127*, 10188–10190.

Johnson M.L., Straume M. (1994). Comments on the analysis of sedimentation equilibrium experiments In T.M. Shuster, T.M. Laue, ed. (Modern Analytical Ultracentrifugation Boston: Birkhauser), pp. 37-65.

Kajander, T., Cortajarena, A.L., Main, E.R.G., Mochrie, S.G.J., and Regan, L. (2005). A New Folding Paradigm for Repeat Proteins. Journal of the American Chemical Society *127*, 10188–10190.

Kobe, B., and Kajava, A.V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends in Biochemical Sciences *25*, 509–515.

Kloss, E., Courtemanche, N., and Barrick, D. (2008). Repeat-protein folding: New insights into origins of cooperativity, stability, and topology. Archives of Biochemistry and Biophysics *469*, 83–99.

Main, E., Lowe, A., Mochrie, S., Jackson, S., and Regan, L. (2005). A recurring theme in protein engineering: the design, stability and folding of repeat proteins. Current Opinion in Structural Biology *15*, 464–471.

Main, E.R.G., Xiong, Y., Cocco, M.J., D'Andrea, L., and Regan, L. (2003). Design of Stable α-Helical Arrays from an Idealized TPR Motif. Structure *11*, 497–508.

Mello, C.C., and Barrick, D. (2004). An experimentally determined protein folding energy landscape. Proc. Natl. Acad. Sci. USA *101*, 14102–14107.

Urvoas, A., Guellouz, A., Valerio-Lepiniec, M., Graille, M., Durand, D., Desravines, D.C., van Tilbeurgh, H., Desmadril, M., and Minard, P. (2010). Design, Production and Molecular Structure of a New Family of Artificial Alpha-helicoidal Repeat Proteins (αRep) Based on Thermostable HEAT-like Repeats. Journal of Molecular Biology *404*, 307–327.

Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K., and Plückthun, A. (2008). Folding and Unfolding Mechanism of Highly Stable Full-Consensus Ankyrin Repeat Proteins. Journal of Molecular Biology *376*, 241–257.

Biographical sketch

Jacob D. Marold was born on December 20th, 1985, in St. Louis Park, MN to Jacqueline Marie Holmbeck and James Daniel Marold. He received his B.S. in Biochemistry and Molecular Biology with academic and departmental honors, and B.A. in Chemistry with academic honors from the University of Minnesota, Duluth in 2009. He joined the Department of Biophysics in the Fall of 2009, and worked with Dr. Doug Barrick on thermodynamics of protein folding cooperativity. He is a member of the Biophysical Society, and the Gibbs Society of Biological Thermodynamics, and will forever be a lifelong academic.